THÈSE

présentée et soutenue publiquement le 23 Juin 2014 par

Amel AISSAOUI

en vue d'obtenir le grade de

Docteur de l'Université des Sciences et Technologies de Lille
(Discipline : Informatique)

Reconnaissance Bimodale de Visages par Fusion de Caractéristiques Visuelles et de Profondeur

---

Composition du jury :

Mme. Sophie TISON	Université Lille 1	Présidente
Mme. Jenny BENOIS-PINEAU	Université Bordeaux 1	Rapporteur
M. Peter VEELAERT	Ghent University	Rapporteur
M. Slimane LARABI	Université Houari Boumediene	Examinateur
M. Chaabane DJERABA	Université Lille 1	Directeur de thèse
M. Jean MARTINET	Université Lille 1	Co-encadrant de thèse

UNIVERSITÉ DES SCIENCES ET TECHNOLOGIES DE LILLE
Laboratoire d'Informatique Fondamentale de Lille | UPRESA 8022
U.F.R. d'I.E.E.A. – Bât. M3 – 59655 VILLENEUVE D'ASCQ CEDEX

Tèl. : +33 (0)3 28 77 85 41 – email : direction@lifl.fr

Remerciements

Tout d'abord, je tiens à remercier mon directeur de thèse, professeur Chaabane Djeraba, de m'avoir accueillie au sein de son équipe pour réaliser ma thèse ainsi que pour ses conseils et son soutien tout au long de ma thèse. Je remercie également Jean Martinet d'avoir co-encadré mon travail, de m'avoir conseillée et guidée tout au long de la thèse.

Je remercie les professeurs Jenny Benois-Pineau et Peter Veelaert d'avoir accepté d'être les rapporteurs de mon mémoire de thèse et les professeurs Sophie Tison et Slimane Larabi d'être examinateurs. Je les remercie de m'avoir fait l'honneur d'accepter de participer à mon jury de thèse.

Je voudrais aussi m'adresser à l'ensemble de l'équipe Fox dans laquelle j'ai réalisé ma thèse. Un grand merci à José et Rémi pour l'ambiance sympathique qu'ils ont su créer au bureau ainsi que pour leur disponibilité et leur aide lorsque j'en avais besoin. Je remercie José aussi pour ses relectures ainsi que pour tous les conseils qu'il a pu me donner. Je remercie Taner pour les nombreuses discussions, et pour l'aide apportée au long de ces années de thèse. Je remercie Pierre pour sa relecture minutieuse de ce mémoire de thèse et pour ses suggestions qui m'ont beaucoup aidé lors de la rédaction. Je remercie Marius, Tarek et Adel pour leurs conseils et pour les diverses discussions que nous avons pu avoir.

Un grand merci à Afifa pour son soutien et pour tous les bons moments qu'on a passés ensemble. Je remercie aussi Donia et Yosra pour leurs encouragements.

J'exprime toute ma gratitude à mes parents qui, malgré la grande distance, étaient toujours là pour m'encourager, me soutenir et m'inciter à avancer. Je remercie aussi mon oncle Nadjib pour les nombreuses discussions qui m'ont beaucoup appris.

Enfin, je remercie tous ceux qui ont contribué de près ou de loin à l'aboutissement de ce travail de recherche.

Résumé

Ce travail s'inscrit dans la thématique de la reconnaissance de visages. Il s'agit de décider de manière automatique de l'identité d'une personne en fonction des traits caractéristiques de son visage. Nous présentons une approche bimodale 2D-3D qui combine des caractéristiques visuelles et de profondeur, afin d'améliorer la précision et la robustesse de la reconnaissance par rapport aux approches monomodales classiques. Dans un premier temps, une méthode d'acquisition 3D par reconstruction stéréoscopique dédiée aux visages est proposée. Cette méthode s'appuie sur un modèle actif de forme permettant de tenir compte de la topologie du visage. Ensuite, un nouveau descripteur DLBP (*Depth Local Binary Patterns*) est défini pour caractériser les informations de profondeur. Ce descripteur étend aux images de profondeur les LBP traditionnels utilisés pour décrire les textures. Enfin, une stratégie de fusion bi-niveaux est proposée, permettant une combinaison à la fois précoce et tardive des deux modalités. Des expérimentations menées sur différentes collections publiques de tests, ainsi que sur une collection spécialement élaborée pour les besoins de l'évaluation, ont permis de valider les contributions proposées dans le cadre de ce travail. En particulier, les résultats ont montré d'une part la qualité des données obtenues à l'aide de la méthode de reconstruction, et d'autre part un gain de précision obtenu en utilisant le descripteur DLBP et la fusion bi-niveaux.

Mots clés : Vision par ordinateur, reconstruction stéréoscopique du visage, reconnaissance bimodale 2D-3D de visages.

Abstract

This work lies in the domain of face recognition. The objective is to automatically decide about a person identity by analyzing his/her facial features. We introduce a 2D-3D bimodal approach that combines visual and depth features in order to provide better recognition accuracy and robustness than classical monomodal approaches. First, a 3D acquisition method dedicated to faces, based on stereoscopic reconstruction, is proposed. It is based on an active shape model to take into account the topology of the face. Then, a novel descriptor named DLBP (*Depth Local Binary Patterns*) is defined in order to characterize the depth information. This descriptor extends to the depth images the traditional LBP originally used for texture description. Finally, a two-stage fusion strategy is proposed, that combines the modalities using both early and late fusions. The experiments conducted with different public datasets, as well as with a new dataset elaborated specifically for the evaluation purposes, allowed to validate the contributions introduced throughout this work. In particular, results have shown the quality of the data obtained using the reconstruction method, and also a gain in precision obtained by using the DLBP descriptor and the two-stage fusion.

Keywords : Computer vision, stereoscopic face reconstruction, 2D-3D bimodal face recognition.

Table des matières

I Introduction, problématique et état de l'art **1**

1 Introduction **3**
1.1 Contexte 4
1.2 Processus général de la reconnaissance de visages 5
1.3 Problématique 5
1.4 Objectif et contributions 8
1.5 Plan de la thèse 9

2 Acquisition de la forme 3D du visage **11**
2.1 Introduction 12
2.2 Techniques d'acquisition active 13
2.3 Techniques d'acquisition passive par reconstruction 3D 14
2.3.1 Les modèles 3D 14
2.3.2 Structure à partir du mouvement (SfM) 16
2.3.3 Structure à partir des ombres portées (SfS) 17
2.3.4 La stéréovision 17
2.4 Reconstruction basée sur la stéréovision 18
2.4.1 Principe général 18
2.4.2 Approches existantes 20
2.4.3 Reconstruction stéréoscopique : le cas du visage 22
2.4.4 Post-traitement de la carte de profondeur du visage 23
2.5 Conclusion et positionnement 25

3 Reconnaissance de visages **27**
3.1 Introduction 28
3.2 Approches de reconnaissance 2D 28
3.2.1 Méthodes globales ou holistiques 29
3.2.2 Méthodes locales 32
3.2.3 Méthodes hybrides 35
3.2.4 Synthèse 36
3.3 Approches de reconnaissance 3D 37
3.3.1 Méthodes d'alignement 37
3.3.2 Méthodes basées sur des propriétés géométriques 38

3.3.3 Méthodes de réduction de dimensionnalité . . . 39
3.3.4 Méthodes basées sur les modèles 3D . . . 40
3.3.5 Synthèse . . . 41
3.4 Approches bimodales 2D-3D . . . 42
3.4.1 Fusion de données brutes . . . 42
3.4.2 Fusion de descripteurs . . . 44
3.4.3 Fusion de décisions . . . 45
3.4.4 Choix de la stratégie de fusion . . . 46
3.4.5 Synthèse . . . 46
3.5 Reconnaissance 3D basée sur les motifs binaires locaux . . . 46
3.5.1 Principe . . . 47
3.5.2 LBP pour la reconnaissance de visages 2D . . . 48
3.5.3 Extensions aux visages 3D . . . 49
3.6 Conclusion et positionnement . . . 54

II Approche proposée pour la reconstruction et la reconnaissance de visages 55

4 Aperçu global de l'approche bimodale 2D-3D de reconnaissance de visages 57
4.1 Introduction . . . 58
4.2 Acquisition . . . 58
4.3 Reconnaissance . . . 59
4.4 Fusion . . . 60
4.5 Conclusion . . . 60

5 Reconstruction 3D du visage 61
5.1 Introduction . . . 62
5.2 Reconstruction stéréoscopique basée sur la structure topologique du visage 62
5.2.1 Construction du modèle de disparité . . . 62
5.2.2 Calcul de la carte de disparité . . . 63
5.3 Post-traitement : débruitage de la carte de profondeur . . . 65
5.3.1 Détection du bruit . . . 66
5.3.2 Suppression du bruit . . . 68
5.4 Conclusion . . . 68

6 Reconnaissance bimodale 2D-3D 71
6.1 Introduction . . . 72
6.2 Extraction des descripteurs 2D et 3D . . . 72
6.3 Descripteur d'images de profondeur . . . 72
6.3.1 Définition du DLBP . . . 73
6.3.2 Stratégie multi-échelles . . . 74
6.3.3 Calcul du seuil . . . 75

6.3.4 Construction des histogrammes 75
6.4 Prédiction de l'identité par fusion 76
6.4.1 Fusion de descripteurs 76
6.4.2 Fusion de décisions 76
6.4.3 Fusion bi-niveaux 78
6.5 Conclusion 78

III Expérimentations, résultats et discussion 81

7 Élaboration d'une collection de tests 83
7.1 Contexte et besoins 84
7.2 Matériel utilisé 84
7.3 Méthodologie 86
7.4 Annotation 87
7.5 Discussion et conclusion 88

8 Évaluation de la méthode de reconstruction stéréoscopique de visages 91
8.1 Introduction 92
8.2 Détection du bruit 92
8.3 Estimation de la profondeur 94
8.3.1 Illustration des résultats de la méthode proposée 95
8.3.2 Comparaison aux méthodes stéréoscopiques 95
8.3.3 Comparaison aux autres méthodes de reconstruction 100
8.4 Conclusion 105

9 Évaluation de la méthode bimodale de reconnaissance de visages 107
9.1 Introduction 108
9.2 Collections de tests 108
9.3 Évaluation du descripteur DLBP 109
9.3.1 Étude des paramètres des DLBP 109
9.3.2 Comparaison avec les autres descripteurs 111
9.4 Évaluation de l'approche globale de reconnaissance bimodale 114
9.5 Conclusion 116

IV Conclusion générale 119

10 Conclusion 121
10.1 Synthèse des contributions 122
10.2 Perspectives 123

Bibliographie 125

Table des figures

1.1 Processus général de la reconnaissance de visages (cas de reconnaissance 2D). 6
1.2 Exemple de variabilité du visage d'une même personne. 7
1.3 Effets du changement des conditions de prise de vue : (a) une prise frontale sous un éclairage ambiant (b) changement de pose (c) changement d'éclairage. 8

2.1 Différentes représentations de la forme 3D du visage. 12
2.2 Reconstruction 3D basée sur 3DMM (source : [22]. 15
2.3 Le modèle de visage *CANDIDE-3* [3]. 16
2.4 Illustration des points correspondants et des disparités sur une image extraite de la collection *Cones* [120]. 19
2.5 Géométrie épipolaire. p_g et p_d sont les deux projections du point p sur les images gauche et droite respectivement. $p1, p2, p3$ sont les points se trouvant dans le même axe de profondeur par rapport à la caméra gauche. Les projections de ces points dans l'image droite se situent dans la même ligne épipolaire. c_g et c_d sont les centres des caméras gauche et droite respectivement. e_g et e_d sont les *épipoles* gauche et droit (les points d'intersection entre les plans images et la ligne reliant les centres des caméras). 19
2.6 Principe de la stéréovision binoculaire. p est un point dans l'espace 3D réel. p_g (resp. p_d) est le point correspondant à p dans l'image gauche (resp. droite). c_g (resp. c_d) est la caméra gauche (resp. droite). f est la focale de la caméra. b est la *baseline*. 20
2.7 Problèmes des méthodes locales d'appariement. O : problème d'ouverture, D : hypothèse implicite (discontinuité de la profondeur). 21
2.8 Problème d'ouverture sur une image de visage. 22
2.9 Artefacts présents dans une carte de profondeur. 24

3.1 Exemple d'application de la méthode *Eigenfaces* sur une collection de 100 visages de 50 personnes. (a) Les 4 premiers visages propres (b) De gauche à droite : images d'une même personne reconstruites à partir de 10, 25, 40 visages propres et enfin l'image originale. 30
3.2 (a) *Eigenfaces* vs (b) *Fisherfaces*. 31

3.3 Distribution de données : (a) les axes correspondants aux composantes principales (PC) et aux composantes indépendantes (IC) (b) le sous-espace avec deux composantes basé sur l'ACP (c) idem avec l'ICA (source : [12]). . . . 32
3.4 Mesures géométriques utilisées par Brunelli *et al.* [29]. 33
3.5 Exemple de représentation d'un visage par un graphe de 9 *jets* avec 4 orientations et 3 échelles (source : [152]). 34
3.6 Deux exemples d'ajustement d'un AAM à une image de test (source : [38]). 36
3.7 Exemple d'alignement de deux visages 3D. 38
3.8 Reconnaissance de visages à l'aide du 3DMM (source : [23]). 41
3.9 Fusion de données brutes (précoce, avant l'extraction de descripteurs). . . . 43
3.10 Fusion de descripteurs (précoce, après l'extraction de descripteurs). 43
3.11 Fusion de décisions (tardive). 44
3.12 Exemple de calcul du code LBP original pour un pixel d'une image (calcul pour le pixel central). 47
3.13 Exemples de voisinages $\mathrm{LBP}_{(R,V)}$ avec différentes valeurs de rayon R et de voisinage V. 48
3.14 Exemple de calcul d'un histogramme de LBP sur une image de visage. . . . 49
3.15 Exemples de formes 3D pouvant être représentées par $LBP_{(1,8)}$. 50
3.16 Confusion entre des formes 3D similaires. (a) Les codes LBP. (b) Les formes 3D possibles correspondantes. 51
3.17 Exemple de calcul des 3DLBP, extrait de [69]. 52
3.18 Exemple des *MS-eLBP-DFs* utilisées pour la reconnaissance 3D par Di Huang *et al.* [65]. 53

4.1 Schéma global de l'approche proposée. Les boîtes grisées représentent nos contributions. 58

5.1 Construction du modèle de disparité. 63
5.2 Décomposition du modèle de disparité. 64
5.3 Différentes projections du modèle de disparité pour un visage sous différentes poses. 64
5.4 Détection des points de découpage et décomposition de la ligne de profondeur. Les points rouges désignent les points de découpage. 67
5.5 Détection des segments bruités. Chaque m_i indique la moyenne du segment s_i. 68
5.6 Débruitage de la carte de profondeur : (a) image bruitée, (b) image des dérivées, (c) détection des segments bruités, (d) image débruitée. 68

6.1 Illustration de l'extraction de c^s et c^m pour un pixel donné, avec $R = 1, V = 8$ et $S^m = 3$. 74
6.2 Exemple d'extraction de $DLBP_{(1,8)}$: (a) image de profondeur (b) matrice des codes de signes M_{c^s} (c) matrice des codes de magnitudes M_{c^m}. 74
6.3 Exemple d'extraction d'histogrammes à partir des deux matrices de signes et de magnitudes : (a) image de profondeur (b) matrice des signes M_{c^s} (c) matrice des magnitudes M_{c^m} (d) histogramme extrait. 76

6.4 Fusion de descripteurs. 77
6.5 Fusion de décisions. 78
6.6 Fusion bi-niveaux. 79

7.1 Système d'acquisition : (a) caméra infrarouge (Kinect) (b) caméra ToF (SR4000) (c) caméra stéréoscopique (Bumblebee XB3). 85
7.2 Exemple de triplet d'images acquises à l'aide du capteur stéréoscopique (Bumblebee XB3). 86
7.3 Exemple des images obtenues à l'aide de la caméra infrarouge (Kinect) : (a) image couleur (b) image de profondeur. 86
7.4 Exemple d'images obtenues à l'aide de la caméra ToF (SR4000) : (a) image infrarouge (b) image de profondeur (c) matrice de confiance (plus le pixel est clair, plus la confiance est bonne). 87
7.5 Les différentes variations acquises pour chaque personne : (a) 3 pour l'éclairage (b) 7 pour l'expression (c) 30 pour la pose. 88
7.6 Annotation d'une image de la base de données. 88

8.1 Débruitage de la carte de profondeur : (a) débruitage global par le filtre médian [15], (b) débruitage local par le filtre médian [98], (c) méthode proposée. 93
8.2 RMS entre les cartes de profondeur débruitées et la vérité terrain. 94
8.3 Cartes de disparité pour des visages sous petites variation de pose. 96
8.4 Cartes de profondeur : (a) vérité terrain (b) *graph cut* (c) *block matching* (d) notre méthode. 97
8.5 Mesure RMS moyenne pour les 118 personnes de la collection Texas. . . . 99
8.6 Mesure PBM moyenne pour les 118 personnes de la collection Texas. . . . 99
8.7 Temps de traitement. 100
8.8 Matrices de similarité (μ est l'erreur moyenne de la diagonale). 101
8.9 Exemples des images utilisées dans les expérimentations. 102
8.10 Coefficients de corrélation pour les 30 premières personnes de Bosphorus. . 102
8.11 Coefficients de corrélation pour les 20 premières personnes de Bosphorus. . 103
8.12 Valeurs de profondeur estimées et réelles pour les 22 points caractéristiques de la personne 1 de la collection Bosphorus. 104
8.13 Écart entre les valeurs de profondeur estimées et réelles pour les 22 points caractéristiques de la personne 1 de la collection Bosphorus. 104

9.1 Impact de la variation du paramètre R sur la précision des DLBP ($H = 25$). 110
9.2 Impact de la variation du paramètre H sur la précision des DLBP. 112
9.3 Comparaison entre l'utilisation des valeurs de profondeur brutes et les DLBP. 113
9.4 Comparaison entre DLBP, LBP et 3DLBP. 114
9.5 Comparaison entre les méthodes monomodales (2D et 3D) et les méthodes bimodales basées sur différentes stratégies de fusion. 117

Liste des tableaux

8.1 Matrice de confusion de la classification des zones de l'image en zones bruitées et non bruitées. 92
8.2 Nombre d'images utilisées par les méthodes de l'état de l'art et par notre méthode. 103
8.3 Coefficients de corrélation pour des visages sous différentes poses. 105
8.4 Moyenne et écart-type pour toutes les images de la collection Bosphorus. . 105

Première partie

Introduction, problématique et état de l'art

Chapitre 1

Introduction

Sommaire

1.1 Contexte 4
1.2 Processus général de la reconnaissance de visages 5
1.3 Problématique 5
1.4 Objectif et contributions 8
1.5 Plan de la thèse 9

1.1 Contexte

Le développement international des communications, tant en volume qu'en diversité (déplacement physique, transactions financières, accès aux services, etc.), implique le besoin de s'assurer de l'identité des individus. Se protéger contre la fraude et l'usurpation d'identité est devenu par conséquent un besoin essentiel au niveau des entreprises, mais aussi au niveau des états. Les gouvernements de nombreux pays cherchent toujours à accroître la sécurité dans les aéroports, aux postes frontaliers et à produire des pièces d'identité plus sûres. Il devient alors nécessaire d'employer des mesures de sécurité efficaces et fiables afin de répondre à ce besoin. La biométrie [1] se présente comme une technologie potentiellement puissante dans le domaine de la sécurité. Les différents moyens biométriques visent à utiliser des caractéristiques spécifiques à chaque personne afin de déterminer ou vérifier l'identité d'une personne. Ces caractéristiques peuvent être physiologiques, comme l'iris, les empreintes digitales, le visage, etc., ou comportementales, comme la signature, la dynamique de l'écriture manuscrite, les gestes, etc. L'avantage de l'utilisation du visage par rapport aux autres modalités biométriques est que la reconnaissance faciale ne nécessite pas théoriquement la coopération des participants. Elle apparaît donc comme une alternative très intéressante, à condition cependant qu'une bonne qualité de reconnaissance soit garantie. Ce caractère non invasif est la raison majeure de l'intérêt grandissant pour la reconnaissance faciale de la part d'organisations de recherche publiques et privées.

En plus des applications liées à la sécurité, la reconnaissance de visages est également utile en interactions homme-machine, réalité virtuelle, indexation et multimédia. Du fait de son vaste champ d'applications, la reconnaissance faciale est devenue un des thèmes de recherche les plus actifs dans le domaine de la vision par ordinateur.

Reconnaître un visage signifie lui affecter une identité parmi un ensemble d'identités connues. En tant qu'êtres humains, nous sommes en mesure de reconnaître et de distinguer les visages sans le moindre effort. Cependant, l'automatisation de cette tâche pour qu'elle soit effectuée par les ordinateurs est une des problématiques les plus communes et les plus complexes de la vision par ordinateur. Dans le milieu des années 1960, des recherches préliminaires cherchant à automatiser le processus de la reconnaissance faciale sont menées [72]. Ces recherches se basent sur la localisation de caractéristiques telles que les yeux, les oreilles, le nez et la bouche sur des photographies pour que le système puisse mesurer les distances et les proportions par rapport à un point de référence, puis les comparer à des données connues. Depuis, la reconnaissance faciale a parcouru un long chemin. Le passage vers des produits commerciaux n'a reçu une impulsion décisive qu'à partir des années 1994-1996, en grande partie grâce à la mise en œuvre d'un programme d'évaluation international, FERET (Face Recognition Technology), organisé par le ministère américain de la défense.

1. La biométrie désigne dans un sens très large l'étude quantitative des êtres vivants. Plus spécifiquement, ce terme est utilisé également dans le sens plus restrictif de l'*identification des personnes* en fonction de caractéristiques biologiques telles que les empreintes digitales ou les traits du visage.

1.2 Processus général de la reconnaissance de visages

La reconnaissance de visages consiste, généralement, en 4 étapes : l'acquisition, le prétraitement, l'extraction de descripteurs, et enfin la classification pour la prise de décision (identité de la personne). L'**acquisition** permet d'obtenir des informations sur le visage de la personne à identifier. Pour cela, différents capteurs peuvent être utilisés. Ils peuvent être classés en deux grandes catégories : capteurs 2D et capteurs 3D. Les capteurs 2D fournissent une image photographique de la personne représentant des informations visuelles [2]. Les capteurs 3D permettent l'acquisition de la forme tridimensionnelle du visage. Selon le type de capteur utilisé, on définit l'approche de reconnaissance comme *reconnaissance 2D* ou *reconnaissance 3D*. Ensuite, une étape de **prétraitement** est souvent appliquée [55]. L'objectif de cette étape est d'abord de déterminer la présence ou non d'un visage dans l'image, ainsi que sa localisation (détection de visages) [140]. Une fois le visage détecté, une étape de normalisation est souvent appliquée. Le visage est normalisé géométriquement dans un but d'alignement. Une normalisation d'illumination peut aussi être effectuée afin de compenser les variations d'éclairage. Après l'étape de prétraitement, l'étape d'**extraction de descripteurs** consiste à extraire des informations importantes à partir de l'image du visage afin de fournir une signature biométrique, ce qu'on appelle *vecteur caractéristique*. Ce dernier doit impérativement satisfaire deux propriétés principales : la discrimination et la robustesse. Un vecteur caractéristique est discriminant s'il prend des valeurs significativement différentes pour les visages de deux personnes distinctes. Quand à la robustesse, il s'agit de l'invariance de ce vecteur à un certain nombre de variabilités (changement d'expression, de pose, d'éclairage, etc.) du même visage. Enfin, la **classification** est effectuée. Elle consiste en deux processus : l'apprentissage et la prédiction. L'apprentissage est effectué sur l'ensemble des vecteurs caractéristiques de visages étiquetés selon leur classe (identité) afin de créer un classifieur. La prédiction consiste à affecter un visage inconnu à une classe en utilisant le classifieur construit par l'étape d'apprentissage. Lors de cette dernière étape, on distingue deux tâches : l'identification, qui permet de connaître l'identité d'une personne, et l'authentification qui permet de vérifier cette identité et confirmer qu'une personne est bien celle qu'elle prétend être. La Figure 1.1 schématise le processus général de reconnaissance 2D de visages.

1.3 Problématique

Malgré toutes les avancées dans le domaine de la reconnaissance de visages, aucun système de reconnaissance fiable n'a encore pu voir le jour. Bien que certains systèmes proposés aient montré une grande efficacité, ils sont souvent limités par les changements de l'environnement, comme le changement d'éclairage, ainsi que par les variations propres aux visages, comme les expressions faciales. Cette variabilité a rendu l'automatisation de la reconnaissance faciale une tâche de grande complexité.

La complexité de la reconnaissance automatique des visages est due à deux difficultés majeures. La première difficulté est la ressemblance structurelle entre les visages. Deux

2. Nous utilisons le terme "image 2D" pour faire référence à ce type d'images.

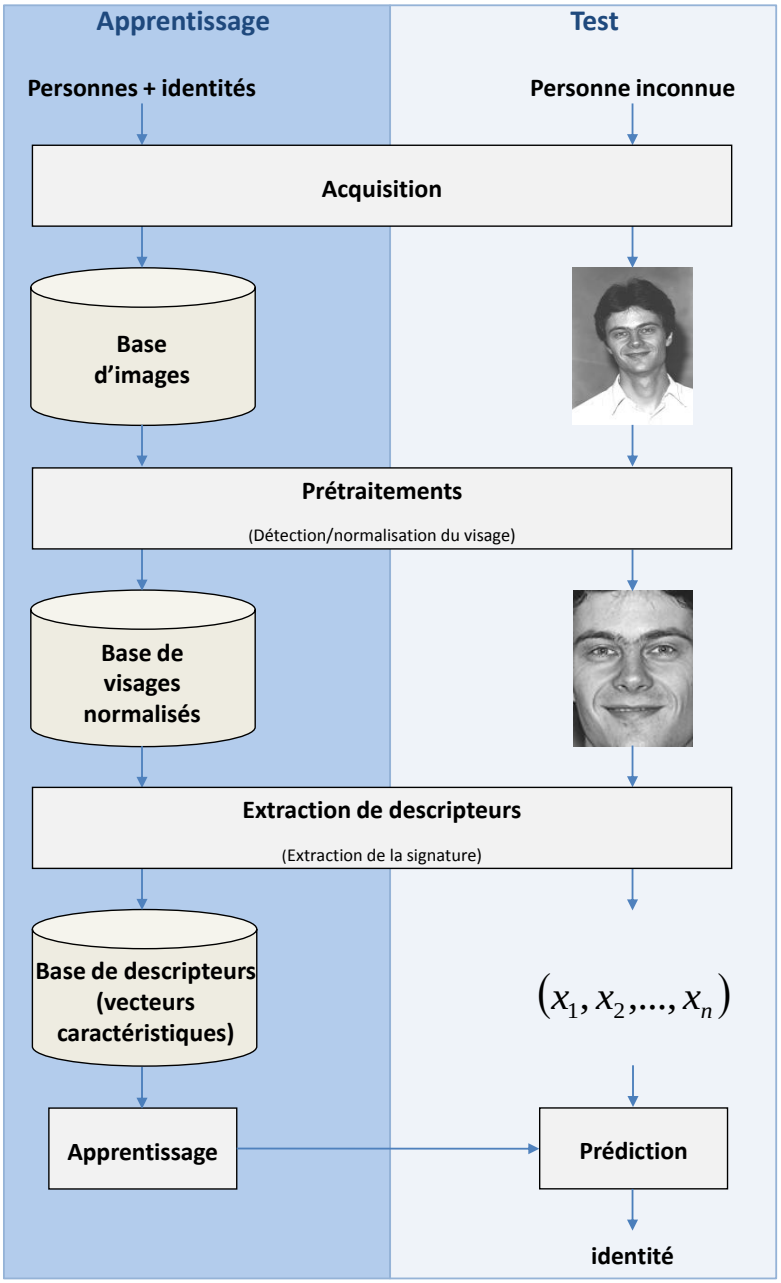

FIGURE 1.1: Processus général de la reconnaissance de visages (cas de reconnaissance 2D).

visages différents sont très proches en termes de structure, car composés des mêmes caractéristiques (yeux, bouche, nez, etc.) dont la localisation et la forme varient légèrement. La deuxième difficulté provient des changements d'apparence d'un même visage acquis dans des conditions d'acquisition différentes (voir la Figure 1.2). Ces différences sont dues, généralement, à des facteurs environnementaux comme les conditions d'éclairage, les caractéristiques des capteurs et aussi leur positionnement par rapport au visage lors de l'acquisition, ou bien aux modifications propres du visage telles que les expressions, les variations de poids

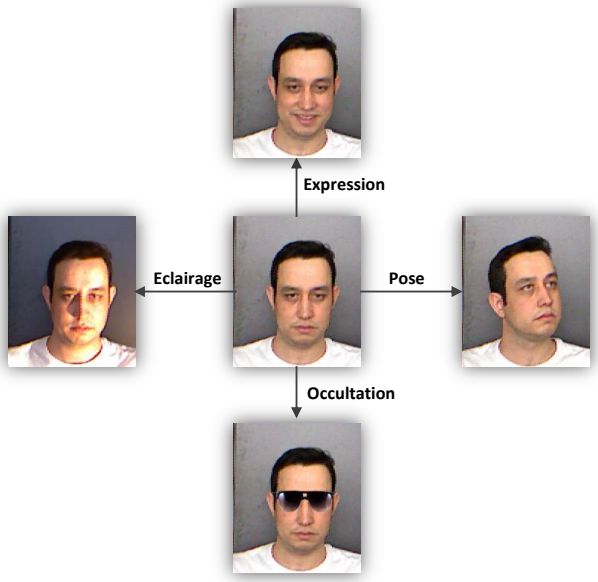

FIGURE 1.2: Exemple de variabilité du visage d'une même personne.

et l'âge. Généralement, il est admis que la différence entre deux visages de personnes différentes acquis dans des conditions identiques est plus faible que celle qui existe pour une même personne dans des conditions différentes. Comme le montre la Figure 1.3, une variation de l'éclairage ou de la pose peut sérieusement altérer l'apparence d'un visage dans l'image : les images de la première ligne appartiennent à la même personne, et les images de la deuxième ligne appartiennent à une autre personne. Pourtant, les images de chaque colonne (a), (b) et (c) semblent plus proches les unes des autres que de leurs correspondants pour lesquelles l'éclairage et la pose ont changé.

Plusieurs méthodes ont été développées pour la reconnaissance 2D de visages [163]. Un certain niveau de maturité est atteint et des taux de reconnaissance très élevés sont reportés [112]. Cependant ces performances sont atteintes dans des environnements contrôlés où les paramètres, tels que l'éclairage, l'angle de prise de vue, la distance de la caméra au sujet, sont maîtrisés. Lorsque les conditions de l'environnement (par exemple : éclairage) ou l'apparence du visage changent (par exemple : pose ou expression faciale), ces performances se dégradent considérablement [112]. Des techniques de reconnaissance 3D ont été proposées comme une solution alternative pour résoudre les problèmes cités ci-dessus [1]. La forme 3D du visage est obtenue en utilisant des équipements de numérisation 3D (par exemple : un scanner laser) ou par des techniques dites de reconstruction. L'avantage des données 3D est qu'elles se caractérisent par leur invariance à la pose et aux conditions d'éclairage. Ceci a permis d'augmenter l'efficacité des systèmes de reconnaissance. L'information de profondeur obtenue par les capteurs 3D représente un aspect différent de l'apparence visuelle obtenue par les capteurs 2D, et peut être considérée comme complémentaire à cette dernière. De plus, les deux modalités sont influencées différemment par les changements environne-

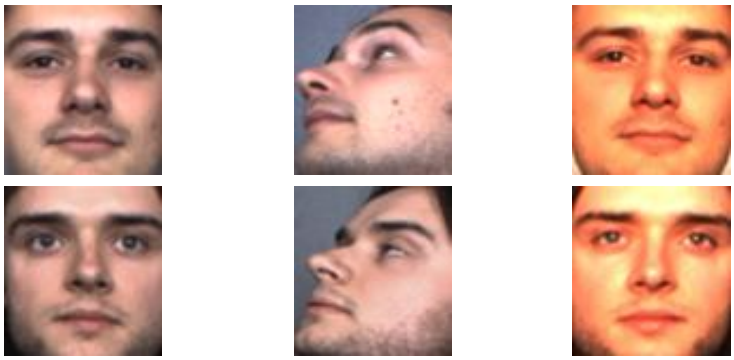

FIGURE 1.3: Effets du changement des conditions de prise de vue : (a) une prise frontale sous un éclairage ambiant (b) changement de pose (c) changement d'éclairage.

mentaux ou par les changements du visage. Alors que la modalité 2D est peu sensible aux changements d'expressions et très sensible aux changements d'éclairage, la modalité 3D est plutôt sensible aux premiers facteurs et invariante aux derniers. Par conséquent, la fusion des deux modalités est susceptible d'augmenter la robustesse et l'efficacité des systèmes de reconnaissance [26]. Ceci a encouragé beaucoup de chercheurs à s'intéresser à la reconnaissance bimodale, qui consiste à combiner des indices obtenus par les deux modalités (2D et 3D) pour la reconnaissance de visages.

1.4 Objectif et contributions

L'objectif de cette thèse est de proposer une approche bimodale 2D-3D de reconnaissance faciale qui combine des mesures d'apparence visuelle et de forme 3D du visage afin d'obtenir un système de reconnaissance discriminant et robuste. La partie 2D de l'approche de reconnaissance proposée se base sur des méthodes de l'état de l'art. En effet, les méthodes de reconnaissance 2D sont nombreuses et les résultats en conditions contrôlées sont satisfaisants, ce domaine de recherche ayant atteint un certain niveau de maturité [163]. Le travail de cette thèse est donc focalisé sur la partie 3D de l'approche, ainsi que sur la question de la fusion bimodale 2D-3D. Afin d'atteindre notre objectif, nous nous intéressons ainsi à trois problématiques : l'acquisition de la forme 3D du visage, la représentation de cette information 3D en extrayant les caractéristiques de forme discriminantes, et enfin, la fusion des données des deux modalités 2D et 3D pour l'identification de visages.

L'information 3D peut être obtenue en utilisant différentes techniques. Les scanners 3D sont les équipements les plus précis et les plus utilisés dans la reconnaissance 3D du visage. Bien que l'utilisation de ces derniers fournisse des données 3D de qualité (nuage de points, maillage), elle n'est applicable que dans un nombre très limité d'applications, principalement à cause de la disponibilité limitée du matériel d'acquisition et son coût élevé, et également

du protocole d'acquisition contraignant – le temps d'acquisition est grand, et enfin parce qu'elle nécessite la pleine coopération de l'individu. Les systèmes d'acquisition stéréoscopiques sont une alternative prometteuse aux scanners 3D. En effet, ces systèmes permettent une acquisition 3D du visage avec un coût moindre et moins de contraintes. Cependant, l'acquisition 3D à l'aide de ces systèmes nécessite une étape de reconstruction à partir des images stéréoscopiques acquises, ce qui constitue une problématique à part entière. L'autre inconvénient est la qualité basse des données 3D que l'on peut obtenir avec ces systèmes par rapport à celle obtenue à l'aide de scanners 3D. Un équipement récent mis au point par Microsoft est le capteur *Kinect*. Ce capteur fournit des données de profondeur d'une qualité moins bonne que les scanners 3D, mais dans un temps très court et ne nécessitant aucune étape de reconstruction. Nous nous intéressons, dans le cadre de cette thèse, aux données provenant des différents types de capteurs. Ceci permet d'étudier l'impact de la qualité des données de profondeur obtenues à l'aide de ces différents équipements pour la reconnaissance de visages. Comme l'information 3D à partir de la stéréoscopie nécessite une étape de reconstruction, nous nous sommes intéressés en premier lieu aux méthodes de reconstruction stéréoscopique et nous avons proposé une **méthode originale de reconstruction spécifique aux visages**. Cela constitue la première contribution de cette thèse. Nous élaborons aussi une collection de test, FBFD [3] (*Fox Bimodal Face Database*) à l'aide de différents capteurs de profondeur afin d'évaluer notre approche.

Une fois l'information 3D obtenue, la deuxième étape de notre approche est l'extraction des caractéristiques. Elle consiste à fournir un vecteur caractéristique représentant la forme 3D du visage. La robustesse et le pouvoir discriminant de ce vecteur sont les deux critères recherchés. Ainsi, la deuxième contribution de cette thèse s'articule autour de l'extraction de mesures discriminantes à partir des données de la forme 3D du visage en définissant un **nouveau descripteur** basé sur les LBP (*Local Binary Pattern*), appelé DLBP (*Depth Local Binary Pattern*).

Enfin, la dernière contribution de ce travail de thèse consiste en la **combinaison des données 2D-3D** dans le processus de la reconnaissance. Nous étudions donc les différentes stratégies de fusion et nous proposons un nouveau schéma de fusion bi-niveaux.

1.5 Plan de la thèse

Ce manuscrit est organisé de la manière suivante. Après l'introduction du contexte de ce mémoire de thèse, le Chapitre 2 présente les travaux existants de l'acquisition de la forme 3D du visage. Nous présentons tout d'abord les différentes techniques d'acquisition 3D existantes en discutant leurs avantages et inconvénients. Ensuite, nous introduisons la technique de stéréovision ainsi que les différentes méthodes proposées dans l'état de l'art pour la reconstruction 3D du visage basée sur cette technique. Dans le Chapitre 3, nous nous intéressons aux méthodes existantes pour la reconnaissance faciale. Nous présentons les différentes catégories de méthodes, notamment 2D, 3D et bimodales 2D-3D, en passant en revue les méthodes les plus connues de chaque catégorie. Ensuite, nous introduisons le descripteur

3. FBFD est disponible à l'URL : http ://www.lifl.fr/FOX/index.php ?page=donnees.

LBP qui est très utilisé pour la reconnaissance faciale 2D en raison de sa simplicité et de son pouvoir de discrimination des visages [5]. Nous présentons ensuite son utilisation en 3D ainsi que les différentes méthodes proposées dans l'état de l'art pour étendre son principe à la 3D. La deuxième partie de ce manuscrit est consacrée à la présentation de l'approche de reconnaissance proposée. Dans le Chapitre 4, un aperçu global de l'approche proposée est présenté. Nous introduisons dans le Chapitre 5, l'approche proposée pour la reconstruction 3D du visage, suivie par l'approche de reconnaissance faciale bimodale dans le Chapitre 6. Dans la troisième partie, nous présentons d'abord la collection de tests élaborée dans le cadre de cette thèse dans le Chapitre 7. Ensuite, une validation expérimentale de l'approche proposée comprenant une évaluation complète quantitative et qualitative de la reconstruction et la reconnaissance de visages est donnée dans les Chapitres 8 et 9 respectivement. Nous concluons ce mémoire de thèse dans le Chapitre 10 en résumant les contributions scientifiques et les résultats obtenus, et nous présentons nos perspectives de recherche.

Chapitre 2

Acquisition de la forme 3D du visage

Sommaire

2.1 Introduction 12
2.2 Techniques d'acquisition active 13
2.3 Techniques d'acquisition passive par reconstruction 3D 14
2.3.1 Les modèles 3D 14
2.3.2 Structure à partir du mouvement (SfM) 16
2.3.3 Structure à partir des ombres portées (SfS) 17
2.3.4 La stéréovision 17
2.4 Reconstruction basée sur la stéréovision 18
2.4.1 Principe général 18
2.4.2 Approches existantes 20
2.4.3 Reconstruction stéréoscopique : le cas du visage 22
2.4.4 Post-traitement de la carte de profondeur du visage 23
2.5 Conclusion et positionnement 25

2.1 Introduction

L'acquisition 3D du visage est un domaine en plein essor dont les applications sont nombreuses comme la réalité virtuelle [134], l'animation de visages [144] et la reconnaissance de visages [92, 124]. Comme nous l'avons présenté dans le chapitre précédent, l'acquisition de l'information 3D du visage est une étape cruciale pour les méthodes de reconnaissance 3D et bimodales. Dans ce chapitre, nous nous intéressons à cette problématique en présentant son principe et les techniques existantes, tout en mettant l'accent sur la reconstruction stéréoscopique qui constitue la première contribution de ce travail de thèse.

Généralement, le modèle 3D du visage est représenté par un maillage 3D ou par une image de profondeur. Le maillage polygonal 3D correspond à une liste de points 3D connectés par des arêtes (polygones) donnant ainsi une représentation de la structure 3D de la surface du visage. Par ailleurs, les images de profondeur correspondent à une représentation 2D de la forme 3D du visage. Chaque point dans le repère (X,Y,Z) est représenté par un pixel (x,y) dans l'image qui stocke sa coordonnée z. Les valeurs sont généralement normalisées afin de donner une image en niveaux de gris allant du blanc au noir (voir la Figure 2.1). Les images de profondeur constituent une bonne alternative aux maillages 3D. En effet, ces images contiennent les informations de forme 3D du visage et elles sont représentées dans un plan 2D, ce qui facilite leur utilisation. De plus, la taille des images de profondeur est petite par rapport aux maillages 3D. En effet, ces derniers sont représentés par des données contenant les coordonnées des points le long des trois axes X, Y et Z, ainsi que les relations entre ces points (les polygones), ce qui nécessite un espace de stockage, et des temps de chargement et de traitement élevés.

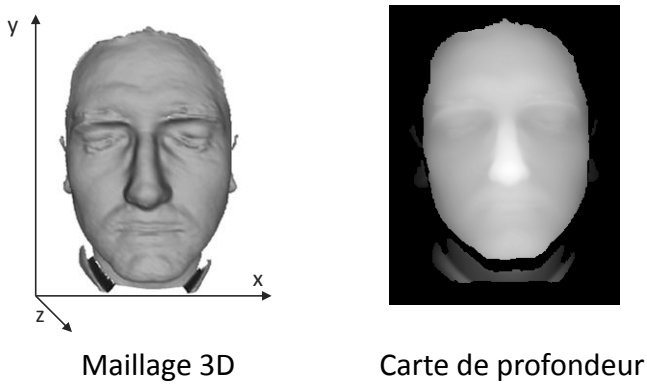

FIGURE 2.1: Différentes représentations de la forme 3D du visage.

Différentes techniques sont utilisées pour acquérir l'information 3D du visage. Nous distinguons deux grandes familles de techniques d'acquisition 3D : les techniques actives et les techniques passives. Dans ce chapitre, nous décrivons les méthodes d'acquisition les plus représentatives en mettant en évidence les méthodes de reconstruction passive en relation avec notre contexte de reconstruction stéréoscopique de visages.

2.2 Techniques d'acquisition active

La technique active est la plus utilisée pour l'acquisition de la forme 3D du visage. Elle consiste à combiner un capteur photographique (ou plusieurs) avec une source lumineuse spécifique contrôlée, afin de mesurer les coordonnées 3D des objets de la scène observée. Différents systèmes d'acquisition active ont été élaborés, nous présentons ici les plus connus.

- **Scanners laser** : le laser consiste en une lumière monochromatique que l'on projette afin d'éclairer une tranche de la scène observée par une caméra associée au dispositif. Ceci permet de calculer la position dans l'espace de la tranche de scène captée en se basant sur le principe de la triangulation active [21]. La triangulation permet de déterminer la position d'un point en mesurant les angles entre ce point et d'autres points de référence (le projecteur de lumière et la caméra) dont la position est connue. Le laser a été beaucoup utilisé pour l'acquisition 3D de visages humains dans des applications différentes. Une collection de visages 3D a été publiée dans le cadre de l'évaluation FRGC (Face Recognition Grand Challenge) [110], comportant des vues frontales des visages et intégrant les deux type de données, 2D et 3D, pour chaque sujet acquis par un scanner laser de type VI900/910 de Konica Minolta. L'objectif du projet est l'évaluation des approches de reconnaissance faciale 2D et 3D sur un corpus d'images de taille significative (un total de 4007 modèles 3D) et riche en variations d'éclairage et d'expressions faciales. Les dispositifs laser présentent de bonnes performances d'acquisition en termes de précision et de résolution. Cependant, quelques difficultés comme les occultations et l'absorption de la lumière par la surface, notamment dans les régions pileuses du visage (sourcils, barbe, moustache) et les régions des yeux, peuvent affecter les résultats de l'acquisition. Par ailleurs, ces équipements sont très onéreux et nécessitent la coopération des personnes qui doivent rester un certain temps face au scanner en bougeant le moins possible afin d'acquérir le modèle 3D. Ceci limite leur utilisation à un nombre restreint d'applications.
- **Systèmes basés sur la lumière structurée** : ce type de technique active consiste à projeter une lumière à motifs structurant sous forme de grille ou des bandes de lumière parallèles. Les motifs sont capturés par la caméra associée. Il s'ensuit une étape d'analyse de la déformation de ces motifs par rapport à leur forme rectiligne d'origine, permettant d'obtenir la géométrie des objets de la scène. Différentes méthodes basées sur la projection de motifs lumineux ont été proposées. Chaque approche propose un schéma de lumière couplé à une méthode de mise en correspondance pour la reconstruction de la forme 3D. Un état de l'art détaillé sur ces méthodes a été présenté dans [115]. Quelques collections de visages ont été construites en se basant sur cette technique pour l'évaluation des méthodes d'analyse de visages 3D, comme la collection 3DRMA [17], contenant des visages 3D de 120 sujets.
- **Systèmes basés sur le temps de vol (*time-of-light, TOF*)** : ils se composent de LED ou de diodes lasers ayant la capacité de générer des impulsions de lumière très rapides [1], et un d'un capteur capable de mesurer le "temps de vol" (de l'ordre de la nanoseconde),

1. L'illumination est généralement dans le proche infrarouge pour ne pas interférer avec la lumière ambiante.

c'est-à-dire le temps que cette impulsion met pour effectuer le trajet aller-retour entre la caméra et l'objet. Le "temps de vol" de cette impulsion est directement proportionnel à la distance entre la caméra et l'objet mesuré. Ceci permet ainsi d'obtenir une image complète de profondeur de l'objet mesuré. Un exemple de ces systèmes est la caméra SR4000 conçue par la société MESA IMAGING.

– **Systèmes basés sur la lumière infrarouge** : le principe de ces systèmes consiste à illuminer une scène par une lumière infrarouge, et ensuite à mesurer la quantité de la lumière incidente réfléchie par les objets. L'hypothèse faite est que plus cette quantité est grande, plus l'objet est proche de la caméra, et inversement. Le capteur Kinect, mis au point par Microsoft, est l'exemple le plus connu de cette catégorie.

2.3 Techniques d'acquisition passive par reconstruction 3D

Les techniques d'acquisition passive opèrent par reconstruction du modèle 3D du visage à partir d'une ou plusieurs images. En vision passive, contrairement à la vision active, une ou plusieurs images prises à partir d'un ou plusieurs angles de vue sont utilisées afin de reconstruire la forme 3D du visage. Au cours des dernières décennies, de nombreuses approches ont été proposées pour l'acquisition passive de la forme 3D du visage (appelée aussi reconstruction 3D), dont celles basées sur les modèles 3D [33], sur le mouvement (*Shape from Motion, SfM*) [81, 128], sur les ombres portées (*Shape from Shading, SfS*) [34, 50], et sur la stéréovision. Nous décrivons par la suite chacune de ces approches.

2.3.1 Les modèles 3D

Ces méthodes consistent à utiliser un modèle 3D afin de reconstruire la forme 3D du visage à partir d'une image 2D. Un modèle 3D, appelé *3D morphable model (3DMM)*, permettant de représenter la forme du visage a été proposé par Blanz et Vetter [22]. Le 3DMM est à l'origine de plusieurs travaux de reconstruction 3D du visage [22, 9, 114, 143, 86]. À partir d'une grande collection de scans 3D de visages alignés, les auteurs ont construit un modèle statistique du visage en termes de forme et de texture. La forme et la texture sont décrites de manière vectorielle, et une ACP est appliquée sur les deux espaces (forme, texture) indépendamment. Les principaux axes de déformation sont caractérisés par les vecteurs propres. De nouveaux visages peuvent donc être décrits par une combinaison linéaire de ces vecteurs, pondérés par un ensemble de paramètres. Ces paramètres peuvent êtres estimés par différentes méthodes d'optimisation [22, 9, 114] cherchant à minimiser la différence entre l'image 2D du visage et l'image synthétisée à partir de son modèle. La Figure 2.2 montre le principe de la méthode proposée par les fondateurs de ce modèle [22].

D'autres chercheurs ont proposé d'utiliser la silhouette du visage au lieu d'une simple image, comme Wang *et al.* [143]. Une silhouette est un contour, une forme ou l'ombre projetée par un objet. Elle fournit des données précises et robustes pour la reconstruction car elle ne dépend que de la forme et de la pose du visage et elle est indépendante de l'éclairage. L'approche proposée par Wang *et al.* [143], ainsi que celle proposée par Lee *et al.* [86] se

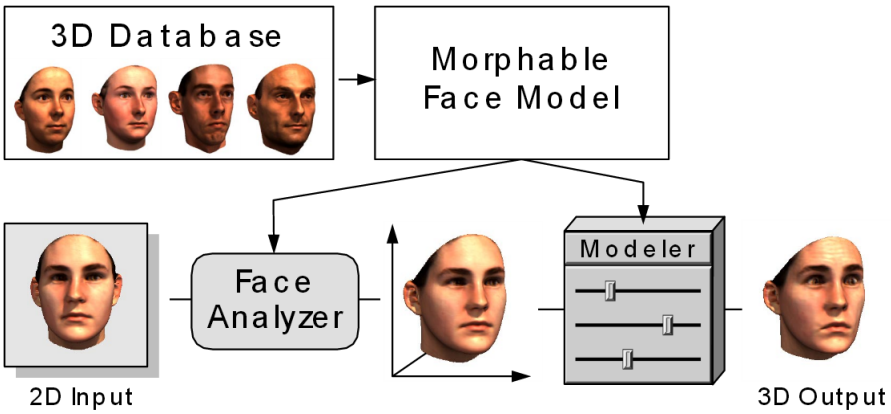

FIGURE 2.2: Reconstruction 3D basée sur 3DMM (source : [22].

basent toutes les deux sur l'utilisation des images de silhouette pour l'étape d'ajustement d'un modèle déformable. L'idée de ces deux approches est similaire à la méthode de Blanz et Vetter [22], sauf qu'au lieu d'utiliser une seule photo, elles se concentrent sur l'acquisition de la géométrie relativement précise d'un visage à partir de plusieurs images de silhouette.

Le problème rencontré par les méthodes basées sur les modèles déformables est que l'algorithme d'optimisation peut converger vers une solution très proche de la valeur initiale, entraînant une reconstruction qui ressemble au modèle générique plutôt qu'au visage particulier qui doit être modélisé. Ainsi, cette méthode peut donner de très bons résultats lorsque le modèle générique présente des similarités significatives avec le visage à reconstruire. Toutefois, si les caractéristiques du modèle générique sont très différentes de celles du visage en cours de reconstruction, le modèle obtenu est susceptible de ne pas représenter fidèlement ce visage. Un autre inconvénient est la nécessité d'une initialisation manuelle afin de faciliter la convergence du système. En effet, en raison des minima locaux de la procédure d'optimisation utilisée pour l'estimation des paramètres du modèle, il est nécessaire d'initialiser cette procédure à proximité de la solution optimale.

Un autre modèle appelé *CANDIDE-3* [3] a été utilisé dans certains travaux de reconstruction de visages en raison de sa simplicité et de sa disponibilité publique. Le modèle *CANDIDE-3* représente la forme d'un visage par un maillage 3D composé de 113 sommets. La formule générale décrivant le modèle *CANDIDE-3* est donnée par deux ensembles : les *Shape Units (SU)* et les *Actions Units (AU)*. Les SU permettent d'adapter le modèle 3D à la physionomie d'une personne. Les AU codent les modifications physiques d'un visage issues de l'activation de muscles faciaux. Elles permettent de reproduire et de s'adapter aux expressions d'un visage. Quelques méthodes sont basées sur ce modèle [81, 127, 128] pour la reconstruction 3D du visage. Cependant, ce modèle est généralement utilisé comme une étape préliminaire d'obtention d'une représentation 3D grossière du visage, et il n'est donc pas adapté à la reconstruction pour l'identification, mais plutôt à d'autres applications comme l'animation du visage.

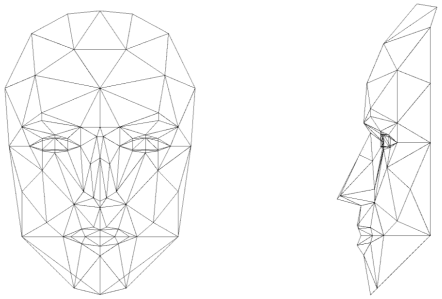

FIGURE 2.3: Le modèle de visage *CANDIDE-3* [3].

2.3.2 Structure à partir du mouvement (SfM)

Cette technique a pour but l'extraction de la forme d'une scène à partir des changements spatiaux et temporels qui se produisent dans une séquence d'images, en exploitant le mouvement relatif entre la caméra et la scène. Le processus de *Shape from Motion (SfM)* consiste en deux étapes : la mise en correspondance entre les images, et l'estimation du mouvement et de la structure. La correspondance entre les images peut être obtenue par des méthodes différentielles [166] ou des méthodes basées sur les primitives d'intérêts (points, lignes, contours, etc.) [149]. La première catégorie de méthodes fournit des mesures denses sur le mouvement apparent (mouvement 2D) en utilisant les dérivées temporelles, et nécessite une séquence d'images séparées par de petits intervalles de temps. Les méthodes de la deuxième catégorie cherchent à établir une correspondance entre les points d'intérêts se trouvant dans des images successives. Ceci peut être fait par des méthodes de corrélation ou bien par des méthodes de suivi comme les filtres de Kalman [77]. La reconstruction et l'estimation du mouvement sont ensuite obtenues par différentes méthodes d'optimisation [95]. La méthode de factorisation, proposée par Kanade [76] a été aussi utilisée pour résoudre ce problème. Pour ce faire, la matrice engendrée par les correspondances calculées entre les images est décomposée en un produit de deux facteurs séparés : la forme et le mouvement.

Le principe de SfM a été appliqué pour la reconstruction de visages dans les travaux de Brand *et al.* [27], Jebara *et al.* [73] et Torresani *et al.* [135]. Par ailleurs, dans [35], Chowdhury *et al.* ont proposé une méthode couplant la technique SfM et un modèle générique du visage. Deux images d'une séquence vidéo fournies en entrée à l'algorithme SfM. La reconstruction 3D obtenue à partir de l'algorithme SfM est fusionnée avec le modèle générique afin de corriger les éventuelles fausses estimations. Les auteurs ont montré que cette combinaison permet d'obtenir une reconstruction plus précise.

Par rapport aux méthodes basées sur les modèles, ces méthodes ne nécessitent pas l'optimisation des paramètres d'un visage 3D moyen. Par conséquent, elles permettent de générer un modèle 3D spécifique à la personne en question. Cependant, la qualité de la reconstruction de visages 3D à partir de deux ou plusieurs images en utilisant la technique SfM est souvent insuffisante. Ceci est lié à l'étape de mise en correspondance entre les images qui est très sensible à la qualité des images, l'homogénéité des surfaces et les occultations.

2.3.3 Structure à partir des ombres portées (SfS)

La technique de *Shape from Shading* (SfS) est une méthode de reconstruction passive qui met en relation les niveaux de gris d'une image et le relief de la scène observée. Elle a été développée par Horn [63] et depuis, de nombreuses approches différentes ont émergé. Cette technique permet d'estimer la forme 3D d'un objet en analysant les variations progressives de l'ombrage dans l'image. Pour expliciter les fondements de SfS, il est nécessaire d'étudier la façon dont les images sont formées. Le niveau de gris d'un pixel dans une image représente la brillance de la scène en ce point. Celle-ci dépend de trois facteurs : l'éclairage de la scène, l'orientation de la surface et ses propriétés de réflectance. Lorsqu'on ne dispose que d'une image de la scène, le niveau de gris lu dans l'image constitue la seule donnée, ce qui rend la résolution du problème très complexe. Afin de réduire le nombre d'inconnues du système, quelques hypothèses sont établies [161] :

- **éclairage** : la scène est éclairée par une source lumineuse unique, émettant un flux lumineux parallèle et uniforme, ou bien par plusieurs sources lumineuses suffisamment éloignées. L'illumination est donc approximativement uniforme sur toute la surface ;
- **réflectance** : la réflectance de la scène, qui décrit la manière dont la lumière est reflétée, est connue. On suppose généralement que la réflectance des surfaces à étudier est homogène et de type diffuse ;
- **modèle de formation** : le modèle de formation de l'image est Lambertien, c'est-à-dire que le niveau de gris d'un pixel de l'image ne dépend que de la direction de la source de lumière et de la normale à la surface ;
- **propriétés de la surface** : la surface est lisse, non texturée et complètement visible (sans occultation).

Le calcul de la normale en tout point de la surface s'effectue alors, grâce à une fonction de minimisation entre la brillance réelle de la scène et la brillance obtenue par estimation de la carte de réflectance.

Plusieurs méthodes ont été proposées pour la reconstruction 3D du visage par la technique SfS [8, 122, 88]. Des a priori sur la forme du visage ont été introduits dans [8, 122] en utilisant un modèle statistique de la forme du visage afin de renforcer le processus SfS utilisé. Aux points d'occultation, les algorithmes SfS ne parviennent pas à déterminer tous les paramètres de la surface. Une façon d'éviter ce problème est d'estimer la profondeur à partir d'images prises à partir de plusieurs points de vue [48].

Bien que le SfS ait montré une efficacité satisfaisante pour la reconstruction 3D, il est basé sur plusieurs hypothèses fortes qui limitent son efficacité. La mise en œuvre de SfS nécessite la connaissance précise des propriétés de réflexion et d'éclairage, et est susceptible de donner de mauvais résultats en raison de certaines hypothèses irréalistes établies sur les propriétés de la surface et l'éclairage.

2.3.4 La stéréovision

L'approche de vision stéréoscopique est une imitation de la stratégie perceptive humaine. Elle permet, par l'utilisation de deux caméras, de fournir la localisation tridimensionnelle des

points de l'espace dont on connaît les projections dans les plans de l'image. Cette technique se base sur une paire stéréoscopique qui consiste en une paire d'images d'une scène donnée, prises sous deux angles d'observation différents. L'objectif principal de la stéréovision consiste à mettre en correspondance, dans une telle paire d'images, les couples de points qui correspondent à un même point de la scène, afin de pouvoir déduire le décalage observé des projections de ces points et ainsi estimer la profondeur. Dans le cadre de cette thèse, nous nous sommes intéressés à cette technique pour générer les données 3D des visages. La section suivante est consacrée à cette technique de reconstruction.

2.4 Reconstruction basée sur la stéréovision

L'extraction de la forme à partir de la stéréovision se réfère à la classe des algorithmes de vision par ordinateur qui appliquent le principe de la stéréoscopie. Ce dernier consiste à déduire les informations de profondeur à partir de deux images prises à partir de différents points de vue [78]. La reconstruction 3D à partir d'images stéréoscopiques repose sur la connaissance a priori de la position et de l'orientation relatives des caméras. En considérant simultanément des images provenant de ces caméras [2], la profondeur de la scène peut être déterminée.

2.4.1 Principe général

Le processus d'estimation de la profondeur dans un système stéréoscopique consiste à estimer ce qu'on appelle la *carte de disparité* de la scène observée. La disparité d'un point est le décalage entre les projections de ce dernier sur les deux images gauche et droite. Pour calculer cette carte, une étape appelée *appariement stéréoscopique* est appliquée. Elle consiste à trouver des pixels correspondants dans les deux images représentant la projection du même point 3D du monde réel. La valeur de cette disparité est proportionnelle à la profondeur du point. En effet, plus l'objet est loin plus sa disparité est petite (voir la Figure 2.4).

Avant d'effectuer l'étape de mise en correspondance, une étape de calibration est généralement appliquée afin de connaître le modèle de projection des caméras ainsi que le déplacement relatif permettant de passer du repère d'une des caméras à l'autre. Les informations de calibration peuvent aussi être utiles pour effectuer la mise en correspondance des pixels. En effet, elles permettent de déterminer la région à laquelle on peut se limiter pour rechercher des points correspondants. On établit ainsi la *contrainte épipolaire*, qui associe à un pixel d'une image une ligne dans l'autre image, le long de laquelle doit se trouver le pixel à apparier dans l'autre image de la paire stéréoscopique. Cette ligne est appelée *ligne épipolaire*. Une étape de rectification, qui consiste à projeter la paire stéréoscopique sur un plan image commun est appliquée. Les lignes épipolaires deviennent donc horizontales.

Une fois la carte de disparité estimée, la profondeur z d'un point $p(x,y,z)$ avec une valeur

2. Il est usuel de se référer à ces caméras par *caméra gauche* et *caméra droite* et aux images correspondantes par *image gauche* et *image droite*.

FIGURE 2.4: Illustration des points correspondants et des disparités sur une image extraite de la collection *Cones* [120].

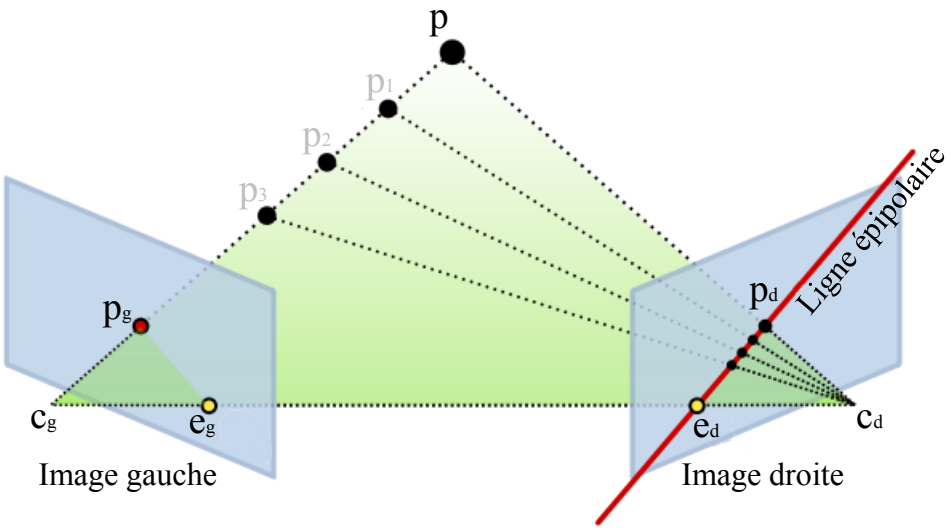

FIGURE 2.5: Géométrie épipolaire. p_g et p_d sont les deux projections du point p sur les images gauche et droite respectivement. $p1, p2, p3$ sont les points se trouvant dans le même axe de profondeur par rapport à la caméra gauche. Les projections de ces points dans l'image droite se situent dans la même ligne épipolaire. c_g et c_d sont les centres des caméras gauche et droite respectivement. e_g et e_d sont les *épipoles* gauche et droit (les points d'intersection entre les plans images et la ligne reliant les centres des caméras).

de disparité d est calculée de la manière suivante :

$$z = \frac{fb}{d} \tag{2.1}$$

où f est la focale de la caméra et b est la *baseline* [3] du système stéréoscopique (voir la

3. Le terme *baseline* désigne la distance entre les deux caméras.

Figure 2.6).

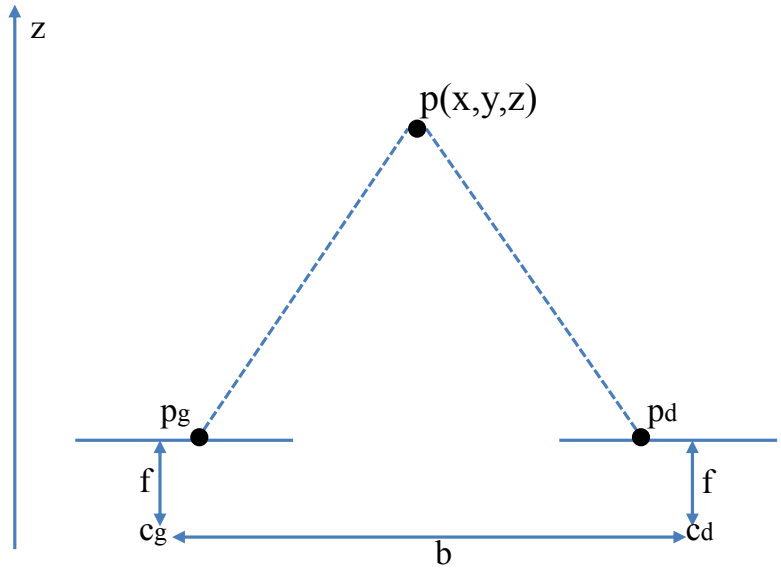

FIGURE 2.6: Principe de la stéréovision binoculaire. p est un point dans l'espace 3D réel. p_g (resp. p_d) est le point correspondant à p dans l'image gauche (resp. droite). c_g (resp. c_d) est la caméra gauche (resp. droite). f est la focale de la caméra. b est la *baseline*.

2.4.2 Approches existantes

Dans l'état de l'art, de nombreux travaux s'intéressant au problème de la reconstruction 3D à l'aide d'un système stéréoscopique sont proposés [119]. Ils peuvent être classés en deux catégories : les méthodes globales et les méthodes locales. Les méthodes globales considèrent l'image dans son intégralité, et cherchent à définir une fonction d'énergie sur l'image. Cette fonction comporte des termes correspondant à des contraintes qu'une mise en correspondance doit satisfaire. Ces conditions concernent généralement le coût de la mise en correspondance et la régularité de la disparité. Le but est alors de minimiser l'énergie en question. Certains paradigmes populaires sont la coupe de graphe (*graph cut*) [80], la propagation de croyances (*belief propagation*) [125] et la programmation dynamique [136]. Les méthodes globales donnent une bonne précision car elles traitent les pixels de manière dépendante. Cependant, leur précision diminue au niveau des pixels éloignés des contours détectés dans l'image, en raison des ambigüités locales au niveau des régions homogènes. Un autre inconvénient des méthodes globales est que les algorithmes d'optimisation utilisés nécessitent un temps de traitement important [80].

Les méthodes locales, souvent appelées méthodes d'appariement de blocs (*block matching*), sont basées sur la corrélation d'intensité. Les traitements requis sont plus légers que pour les méthodes globales et peuvent ainsi être utilisées dans des applications en temps réel.

Les algorithmes d'appariement stéréoscopique basés sur la corrélation produisent généralement des cartes de profondeur denses en calculant les coûts de mise en correspondance de chaque pixel selon toutes les valeurs de disparité d'un intervalle donné. Un seuil de confiance est parfois utilisé afin de définir la valeur maximale du coût, en deçà de laquelle la mise en correspondance n'est pas considérée. Ensuite, ces coûts peuvent être agrégés au sein d'une certaine fenêtre de voisinage. Enfin, l'algorithme recherche la correspondance qui donne le coût le plus bas pour chaque pixel. Différentes mesures de similarité sont utilisées dans les méthodes basées sur la corrélation. Les plus courantes sont les suivantes : la somme des différences absolues (*Sum of Absolute Differences, SAD*), la somme de différences au carré (*Sum of Squared Differences, SSD*), la corrélation croisée normalisée (*Normalized Cross Correlation, NCC*) et la somme des distances de Hamming (*Sum of Hamming Distances, SHD*).

Les méthodes locales d'appariement de blocs souffrent de deux problèmes essentiels (voir la Figure 2.7) :

- **hypothèse implicite** : tous les points appartenant à une même fenêtre de corrélation font l'objet d'une disparité unique, ce qui est incorrect au niveau des bords des objets (discontinuité de la profondeur) ;
- **problème d'ouverture** : les informations discriminantes peuvent être trop faibles dans certaines régions, notamment dans les régions homogènes.

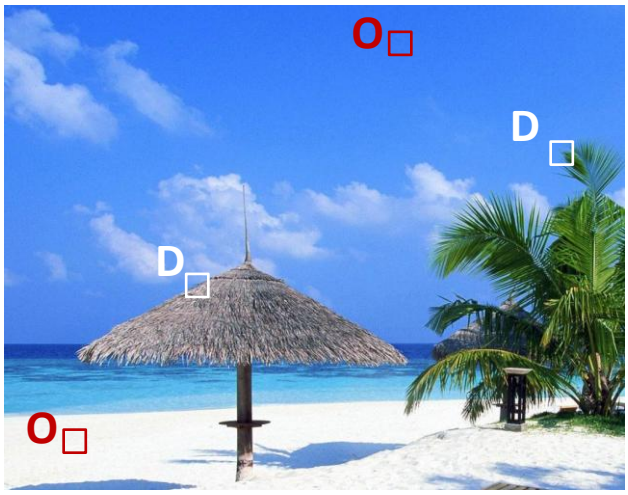

FIGURE 2.7: Problèmes des méthodes locales d'appariement. O : problème d'ouverture, D : hypothèse implicite (discontinuité de la profondeur).

Le fondement des approches de reconstruction stéréoscopique réside dans la définition d'un schéma de mise en correspondance stéréoscopique pour une paire d'images donnée. Ce point est particulièrement crucial avec des images de visages car ces dernières sont très peu texturées. Cela conduit à de nombreuses ambiguïtés pour les algorithmes de mise en correspondance. Par conséquent, peu d'approches de reconstruction 3D dense du visage à partir de la stéréovision ont été proposées.

2.4.3 Reconstruction stéréoscopique : le cas du visage

Le processus de reconstruction stéréoscopique dans le cas du visage est très complexe quelle que soit l'approche, locale ou globale, en raison de l'homogénéité des régions du visage et donc du problème d'ouverture (voir la Figure 2.8). Cette caractéristique représente un sérieux problème car l'information d'intensité au niveau des zones homogènes n'est pas suffisamment discriminante pour un appariement précis. Par conséquent, il est difficile de trouver le pixel correspondant car tous les pixels donnent des valeurs de similarité très proches. Des incertitudes et des erreurs de mise en correspondance causées par cette homogénéité donnent lieu à la présence de bruit dans la carte de profondeur estimée.

FIGURE 2.8: Problème d'ouverture sur une image de visage.

Afin de reconstruire la forme 3D d'un visage, des méthodes ont été proposées dont l'objectif est de combiner d'autres techniques de reconstruction à la technique stéréoscopique afin de la rendre plus robuste aux zones homogènes. La plupart des méthodes proposées sont basées sur une étape d'ajustement de la carte de profondeur estimée à un modèle 3D générique [85, 96, 107, 164]. Dans [85], Le *et al.* proposent d'effectuer une estimation éparse des points de la forme du visage basée sur des points clés 3D, et ensuite utiliser un modèle linéaire déformable pour obtenir la forme et la texture dense. Mallick *et al.* [96] utilisent un ensemble de correspondances entre des points d'intérêts sélectionnés manuellement sur la paire stéréoscopique pour calculer une matrice de translation et de rotation. Ces matrices sont utilisées pour ajuster le modèle 3D au nuage des points épars sélectionnés. De la même manière, Park et Jain [107] ont généré une reconstruction éparse du visage à partir d'un ensemble de points sélectionnés manuellement sur la paire stéréoscopique qui a été ensuite alignée à un modèle générique en utilisant la méthode *Thin-Plate Spline (TPS)*. Dans [164], Zheng *et al.* ont utilisé un visage 3D de référence comme intermédiaire pour l'estimation de la correspondance. Des images stéréoscopiques avec leurs correspondances connues a priori sont d'abord synthétisées à partir du modèle de référence. Ces correspondances sont ensuite

étendues aux images stéréoscopiques entrantes en utilisant l'alignement et la déformation du modèle 3D de référence. Le modèle 3D du visage peut donc être reconstitué à partir des images stéréoscopiques de manière fiable.

Le problème majeur de ces méthodes est le coût élevé du temps de traitement lié à l'étape d'ajustement, ainsi que le besoin d'une initialisation manuelle dans certains cas [164]. Un autre inconvénient de ces méthodes est que les visages qui en résultent sont plus semblables au modèle générique qu'à leurs propres modèles. Dans [90], Lengagne *et al.* ont proposé une approche itérative pour déformer un modèle de maillage 3D à partir de deux images stéréoscopiques. Ils ont incorporé des contraintes géométriques liées aux propriétés différentielles de la surface dans leur approche. Cela permet d'améliorer la correspondance dans les zones de la surface du visage qui posent problème en raison de l'ambigüité. Cette méthode ajoute un coût de calcul additionnel à cause de l'étape de déformation itérative et du calcul des courbes principales pour chaque point du visage. De plus, elle est très sensible au bruit car elle utilise la dérivée seconde pour le calcul de ces courbes.

Quelques tentatives utilisant la méthode SfS (voir la Section 2.3.3) ont été proposées pour améliorer le processus d'appariement stéréoscopique. Dans [41], Cryer *et al.* ont proposé de fusionner la carte de profondeur estimée en utilisant l'appariement stéréoscopique avec celle obtenue par la technique SfS. Le processus de fusion est basé sur l'hypothèse que le SfS fonctionne bien sur les zones homogènes. Ceci est considéré comme complémentaire à l'appariement stéréoscopique qui donne de bons résultats dans les zones texturées. Plusieurs méthodes utilisant différentes versions des techniques SfS pour améliorer les résultats de l'appariement stéréoscopique ont été proposées [34]. Cependant, dans ces méthodes, chaque processus (stéréovision et SfS) est sensible aux conditions d'éclairage. Aussi, dans les algorithmes de SfS, il est courant de considérer que la matrice de réflectance de la surface en question est donnée ou que sa forme est connue.

Bien que plusieurs solutions aient été proposées pour la reconstruction 3D à partir de la stéréovision, acquérir efficacement des informations de profondeur du visage à partir d'images stéréoscopique reste toujours un problème difficile à cause du problème d'ouverture, qui se pose dans le cas de visages avec davantage d'acuité.

2.4.4 Post-traitement de la carte de profondeur du visage

La carte de profondeur estimée contient généralement des artefacts produits par des erreurs de mise en correspondance et des incertitudes. Par conséquent, il est souvent nécessaire de recourir à une étape de post-traitement afin de les éliminer. Deux types d'artefacts sont possibles dans la carte de profondeur :

- valeurs manquantes : aucun pixel correspondant n'est trouvé. La disparité du pixel reste donc indéfinie et la valeur de profondeur ne peut pas être calculée. Ces valeurs sont représentées par des *trous* dans la carte de profondeur ;
- valeurs erronées : la position du pixel correspondant est erronée. La disparité estimée est donc différente à la disparité réelle (supérieure ou inférieure). Un bruit est donc

présent dans la carte de profondeur [4].

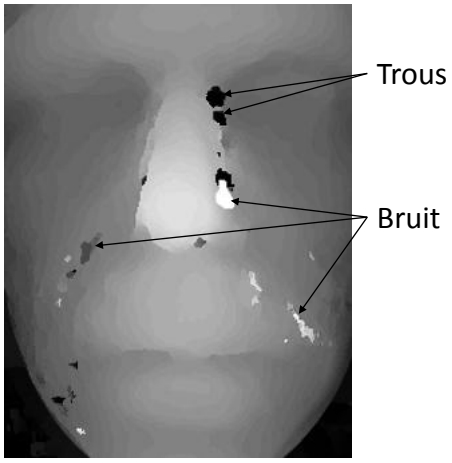

FIGURE 2.9: Artefacts présents dans une carte de profondeur.

Dans l'état de l'art, différentes méthodes ont été proposées pour le débruitage de la carte de profondeur. Afin de traiter les données manquantes, il s'agit de trouver le paramétrage approprié qui permet la reconstruction de ces données à l'aide des données disponibles (voisinage). Ceci est généralement réalisé par un algorithme d'interpolation linéaire ou cubique [74, 15, 66, 147, 47]. Le traitement des données erronées consiste souvent en l'application des filtres de réduction de bruit sur la carte de profondeur. Le filtre médian est couramment appliqué afin de débruiter et lisser les valeurs de profondeur [74, 15, 66, 47]. Dans [147], Wang *et al.* ont utilisé trois filtres gaussiens avec des variances différentes pour enlever les pics, remplir les petits trous et lisser les données. L'avantage de ces méthodes est qu'elles peuvent réduire le bruit de différentes tailles en adaptant les paramètres du filtre. Cependant, elles traitent la carte de profondeur globalement, et peuvent ainsi causer la perte des données exactes, car elles affectent toute l'image (incluant les pixels avec une valeur de profondeur correcte). Afin de pallier ce problème, une approche consiste à d'abord à identifier les valeurs erronées et ensuite à les corriger de la même manière que les données manquantes. Afin de détecter ces valeurs erronées, une méthode consiste à parcourir la carte de profondeur et à appliquer un seuillage sur la valeur absolue entre la valeur de profondeur d'un pixel et la valeur médiane de son voisinage. Seulement quand ce seuil est dépassé, le pixel est identifié comme un bruit [98]. La détection du bruit permet un débruitage plus précis de la carte de profondeur et préserve ainsi les données correctes. Cependant, si la zone bruitée est grande, le bruit devient plus difficile à détecter. En effet, dans ce cas de figure, l'information de profondeur du voisinage est également bruitée.

4. Un bruit peut correspondre à des pics positifs ou des pics négatifs (fossés).

2.5 Conclusion et positionnement

Dans ce chapitre, nous avons présenté différentes techniques d'acquisition de la forme 3D du visage. La problématique d'estimation de la forme 3D du visage à partir de données stéréoscopiques est détaillée, ainsi qu'un nombre de travaux représentatifs de cette approche. Enfin, un aperçu général sur les méthodes de post-traitement des cartes de profondeur de visages est présenté.

L'aspect homogène des régions du visage constitue un problème majeur lors du processus d'appariement stéréoscopique. Différentes solutions ont été proposées afin de pallier ce problème. Elles sont basées essentiellement sur l'utilisation de modèles 3D du visage et sur la technique SfS. Cependant, l'utilisation de modèles 3D nécessite des méthodes d'optimisation souvent coûteuses en temps de calcul, et susceptibles de converger vers un minimum local. La technique SfS est complémentaire à l'appariement stéréoscopique car elle peut être appliquée sur des zones homogènes. Cependant, elle est basée sur plusieurs hypothèses non-valides dans les cas réels. Dans le cadre de notre travail, nous nous intéressons à ce problème et nous proposons une méthode de reconstruction stéréoscopique qui ne nécessite ni modèle générique 3D, ni recours à d'autres techniques comme SfS. L'idée consiste à incorporer des a priori concernant la topologie des visages, ainsi que le caractère lisse de leurs formes pour guider le processus d'appariement.

Dans la deuxième partie de ce chapitre, nous avons évoqué la problématique de débruitage des cartes de profondeur reconstruites. Nous avons vu que, la majorité des méthodes de débruitage est basée sur les filtres de réduction de bruit et notamment sur le filtre médian. L'application du processus de débruitage sur la totalité de la carte de profondeur peut affecter toutes les données de profondeur. Ainsi, il est préférable de traiter uniquement les zones bruitées afin d'éviter la perte des données correctes. Pour cela, les données erronées doivent d'abord être détectées et ensuite traitées. Cependant, ceci devient problématique lorsque la zone bruité est grande. Afin de résoudre ce problème, nous proposons une méthode pour le post-traitement des cartes de profondeur des visages basée sur la caractéristique lisse du visage et sur une analyse statistique à la fois locale et globale.

Chapitre 3

Reconnaissance de visages

Sommaire

3.1 Introduction 28
3.2 Approches de reconnaissance 2D 28
3.2.1 Méthodes globales ou holistiques 29
3.2.2 Méthodes locales 32
3.2.3 Méthodes hybrides 35
3.2.4 Synthèse 36
3.3 Approches de reconnaissance 3D 37
3.3.1 Méthodes d'alignement 37
3.3.2 Méthodes basées sur des propriétés géométriques 38
3.3.3 Méthodes de réduction de dimensionnalité 39
3.3.4 Méthodes basées sur les modèles 3D 40
3.3.5 Synthèse 41
3.4 Approches bimodales 2D-3D 42
3.4.1 Fusion de données brutes 42
3.4.2 Fusion de descripteurs 44
3.4.3 Fusion de décisions 45
3.4.4 Choix de la stratégie de fusion 46
3.4.5 Synthèse 46
3.5 Reconnaissance 3D basée sur les motifs binaires locaux 46
3.5.1 Principe 47
3.5.2 LBP pour la reconnaissance de visages 2D 48
3.5.3 Extensions aux visages 3D 49
3.6 Conclusion et positionnement 54

3.1 Introduction

La reconnaissance automatique du visage est une des techniques biométriques les plus communes et populaires. Grâce à son caractère non-intrusif ainsi que ses nombreux domaines d'application tels que la sécurité, la communication et le loisir, la reconnaissance faciale a beaucoup attiré l'attention des chercheurs depuis quelques décennies. Beaucoup de travaux de recherche en reconnaissance automatique de visages ont été menés, dans lesquels des connaissances dans les domaines de la reconnaissance des formes, du traitement d'images et des statistiques ont été appliquées. Cependant, la reconnaissance efficace de visages reste encore un grand défi pour les chercheurs en vision par ordinateur et reconnaissance de formes. De nombreuses méthodes de reconnaissance de visages ont été proposées au cours de ces dernières années. Le but de ce chapitre est de donner un panorama des méthodes les plus représentatives. Une classification des méthodes de reconnaissance selon le type des données en entrée du système est adoptée. On distingue trois grandes classes de méthodes de reconnaissance de visages :

- **la reconnaissance 2D**, qui traite principalement des images représentant l'apparence visuelle des visages. Ces images, en couleurs ou en niveaux de gris, sont généralement acquises via un appareil photographique ordinaire ;
- **la reconnaissance 3D**, qui exploite des données concernant la forme 3D des visages. Ces données peuvent être obtenues à l'aide d'équipements spéciaux, comme par exemple un scanner laser ;
- **la reconnaissance bimodale 2D-3D**, qui se fonde sur les deux modalités pour représenter le visage. Le but est d'exploiter la complémentarité des données 2D et 3D.

Dans la suite de ce chapitre, quelques méthodes représentatives de chaque catégorie, ainsi qu'une analyse de leurs avantages et inconvénients, sont présentées.

3.2 Approches de reconnaissance 2D

Les approches de reconnaissance 2D ont été les premières étudiées dans la littérature, et sont le sujet de la majorité des travaux de recherches dans le domaine de la reconnaissance de visage. Dans l'état de l'art, et selon la méthode retenue pour représenter les visages, trois catégories d'approches ressortent [163] :

- **les méthodes globales**, qui traitent le visage dans son ensemble comme une seule entité. Elles considèrent donc la région entière du visage et ne font pas la différence entre les différents composants du visage ;
- **les méthodes locales**, qui représentent un visage par un ensemble de régions sur lesquelles des statistiques sont calculées afin de créer une description du visage ;
- **les méthodes hybrides**, qui, comme le système de perception humain, utilisent à la fois des caractéristiques locales et globales de la région du visage afin de l'identifier [163]. Cela permet d'obtenir plus de précision et de combiner les avantages des deux méthodes.

Dans chacune de ces catégories, une multitude de méthodes de classification sont considérées. Parmi les plus utilisées, on peux citer la méthode des k-plus proches voisins (*K-Nearest Neighbors, KNN*) [139], les réseaux de neurones artificiels (*Artificial Neural Networks, ANN*) [57], les machines à vecteurs de support (*Support Vector Machines, SVM*) [25], le boosting (comme *Adaptative Boosting, AdaBoost*) [51], etc. Le choix de ces méthodes est principalement guidé par l'application finale, le type et la quantité de données, la dimensionnalité de l'espace de description, et les contraintes techniques.

3.2.1 Méthodes globales ou holistiques

Le principe de ces méthodes est d'utiliser l'image du visage dans sa globalité. Cette représentation basée sur l'apparence globale a l'avantage de conserver implicitement toute l'information de texture et de forme utile pour différencier les visages. La majorité des méthodes appartenant à cette catégorie sont des méthodes statistiques de réduction de la dimension de l'espace de description. Les visages sont représentés par des vecteurs de grande dimension obtenus en concaténant les valeurs de tous les pixels de l'image du visage. Ensuite, ces vecteurs sont projetés dans un nouvel espace de dimension inférieure pour maximiser la variance des données. Une image de visage peut ainsi être représentée par une combinaison linéaire des vecteurs de la base du nouvel espace, appelés vecteurs propres. Une méthode de classification souvent utilisée dans cette approche est la méthode des k-plus proches voisins. Plusieurs mesures de similarité basées sur différentes distances comme la distance de Manhattan, le cosinus, la distance de Mahalanobis [19], ainsi qu'une combinaison de toutes ces mesures [43] ont été ensuite utilisées. Selon les résultats avancés par Beveridge *et al.* [18], la distance de Mahalanobis-cosinus est la plus précise dans ce contexte.

La méthode la plus ancienne dans l'état de l'art est celle des visages propres (*Eigenfaces*), proposée par Turk et Pentland [139]. Elle est basée sur l'Analyse en Composantes Principales (ACP ou *Principal Component Analysis, PCA*), utilisée initialement par Sirovich et Kirby [121] pour la compression de visages. Les *Eigenfaces* sont les composantes principales de l'espace des visages. Une image de visage peut ainsi être représentée par une combinaison linéaire des Eigenfaces (voir la Figure 3.1).

L'approche *Eigenfaces* est une étape importante vers la reconnaissance basée sur l'apparence globale des visages et elle est encore souvent considérée comme une méthode de comparaison de base pour démontrer la performance minimale attendue d'un tel système. L'avantage de cette méthode est son invariance au bruit. Il est intéressant de noter que l'aspect global des images reconstruites par l'ACP est meilleur que celui des images originales. De plus, la méthode est moins sensible aux petites occultations, aux changements de fond et au flou des images [163]. Cependant, la méthode est très sensible à la variation de l'éclairage et de la pose, du fait qu'elle utilise la valeur directe de tous les pixels de l'image. Par conséquent, plusieurs extensions ont été proposées pour pallier ces problèmes. Une extension appelée *view-based Eigenfaces* a été proposée par Pentland *et al.* [109] afin de résoudre le problème de variation de pose. Dans cette extension, un espace propre est construit pour chaque pose. L'espace propre qui décrit le mieux une image d'entrée est sélectionné en calculant l'erreur de description à l'aide des vecteurs propres de chaque espace. L'image est

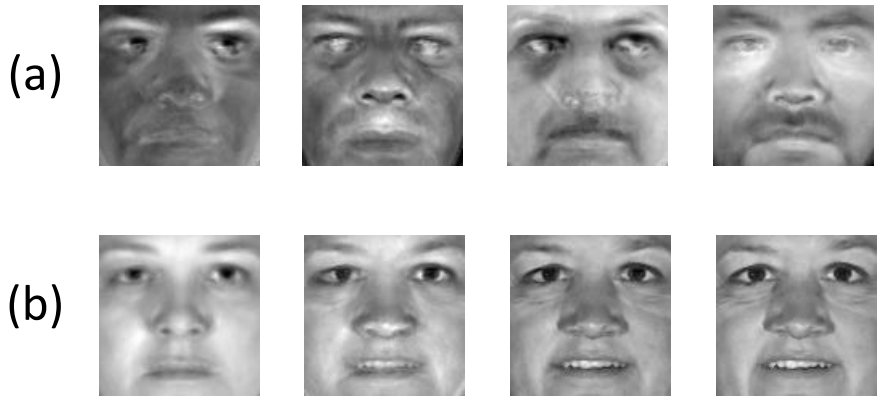

FIGURE 3.1: Exemple d'application de la méthode *Eigenfaces* sur une collection de 100 visages de 50 personnes. (a) Les 4 premiers visages propres (b) De gauche à droite : images d'une même personne reconstruites à partir de 10, 25, 40 visages propres et enfin l'image originale.

ensuite décrite en utilisant les vecteurs propres de l'espace sélectionné. La limite principale de cette méthode est que plusieurs images de visages sous différentes poses connues sont nécessaires pour construire les espaces propres.

L'utilisation de l'analyse en composantes principales engendre des directions de projections qui maximisent la dispersion totale de tous les visages. Ceci mène à la prise en compte des variations indésirables telles que les changements d'éclairage et les différentes expressions des visages. Par conséquent, les distances entre les images d'un même visage sous différentes vues (changements d'éclairage, de pose, etc.) sont plus grandes que celles entre deux visages différents acquis dans les mêmes conditions, puisque les variations de l'apparence visuelle dans ce cas sont plus petites (voir la Figure 1.3 dans l'introduction à la Section 1.3). Ceci explique que la projection par ACP est optimale pour représenter les visages d'une manière compacte, mais moins efficace d'un point de vue discrimination. Lorsque l'objectif est la classification plutôt que la représentation, le calcul de la distance entre deux visages dans l'espace généré par l'ACP peut ne pas donner les résultats les plus pertinents. Dans de tels cas, on souhaite trouver un sous-espace dans lequel les images d'une même classe sont situées en un seul endroit et les différentes classes sont aussi éloignées les unes des autres que possible. Une technique pour atteindre cet objectif est l'Analyse Discriminante Linéaire (ADL ou *Linear discriminant analysis, LDA*), introduite en 1936 par Robert Fisher [49]. Cette technique a beaucoup été utilisée dans le domaine de la reconnaissance de visages. À la différence de l'ACP, l'ADL est une méthode supervisée qui cherche à maximiser les variances entre les classes, tout en minimisant la variance à l'intérieur des classes, afin de trouver les directions de projections les plus discriminantes dans l'espace propre. Pour cela, l'objectif de la technique ADL est de maximiser le rapport des variances inter-classes sur les variances intra-classes. Belhumeur *et al.* [13] ont proposé une méthode appelée *Fisherfaces* qui consiste à appliquer une ACP sur l'ensemble des images afin de réduire l'espace d'origine, puis une ADL afin de calculer les *Fisherfaces*. La réduction d'es-

pace par l'ACP proposée par les auteurs résout un problème dit de *singularité* de la matrice calculée par l'ADL, du fait que les données sont souvent sous-représentées (c'est-à-dire que la taille des vecteurs images d'une classe donnée est très supérieure à leur nombre). Dans la Figure 3.2, on remarque que, visuellement, les *Fisherfaces* sont moins représentatives que les *Eigenfaces*. Ceci est dû au fait que le but de l'ACP est la représentation des visages alors que l'ADL considère essentiellement la séparabilité inter-classes. Les performances des systèmes de reconnaissance basés sur l'ADL sont souvent meilleures que celles basées sur l'ACP. Cependant, les méthodes basées sur l'ADL ne fonctionnent bien que lorsque beaucoup d'images par personne sont fournies. Dans [13], Belhumeur *et al.* ont montré que dans le cas où peu d'images sont disponibles par personne, la capacité de généralisation des vecteurs discriminants devient très faible. Dans de tels cas, l'ACP est susceptible de donner des résultats meilleurs.

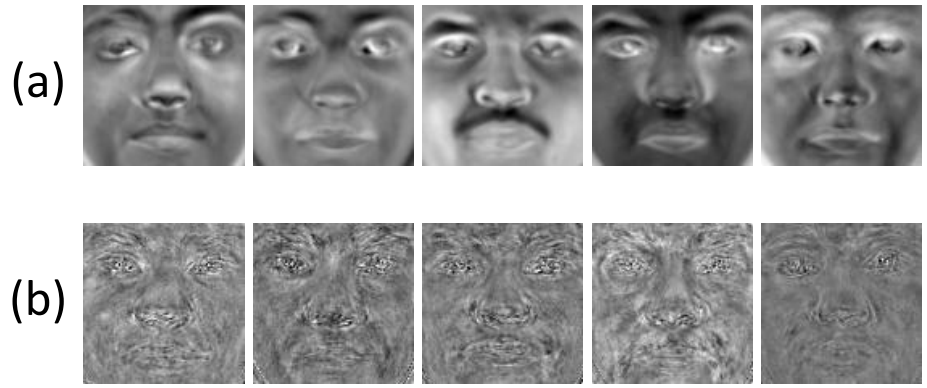

FIGURE 3.2: (a) *Eigenfaces* vs (b) *Fisherfaces*.

Une autre méthode d'analyse de données qui a été largement utilisée dans la reconnaissance de visage est l'Analyse en Composantes Indépendantes (ACI ou *Independent Composent Analysis, ICA*). L'ACI est historiquement connue comme une méthode de séparation aveugle de sources, et a été appliquée pour la reconnaissance faciale. L'objectif de l'ACI est de minimiser la dépendance statistique des projections. À la différence de l'ACP qui décorrèle les données en minimisant la dépendance de second ordre (matrice de covariance), l'ACI effectue une projection linéaire en minimisant la dépendance d'ordre supérieur entre les données, afin de les rendre statistiquement indépendantes. Cette méthode a été initialement appliquée dans le cadre de la reconnaissance de visages par Bartlett *et al.* [11, 12]. La Figure 3.3, tirée de [12], montre les projections d'un ensemble de données 3D par l'ACP et l'ACI. Les auteurs ont rapporté une amélioration notable de performance par rapport à la méthode *Eigenfaces*.

D'autres extensions non-linéaires des méthodes présentées ci-dessus ont aussi été proposées dans le domaine de la reconnaissance de visages, comme la méthode *Kernel Eigenfaces* proposée par Yang *et al.* [159]. L'ADL a aussi été étendue pour le cas non-linéaire [100] où des noyaux ont été utilisés afin de transformer l'espace linéaire d'entrée en un espace non-linéaire dans lequel l'ADL est effectuée pour trouver une séparation linéaire des visages. Une version non-linéaire de l'ACI a été utilisée dans [97] pour la reconnaissance de visages.

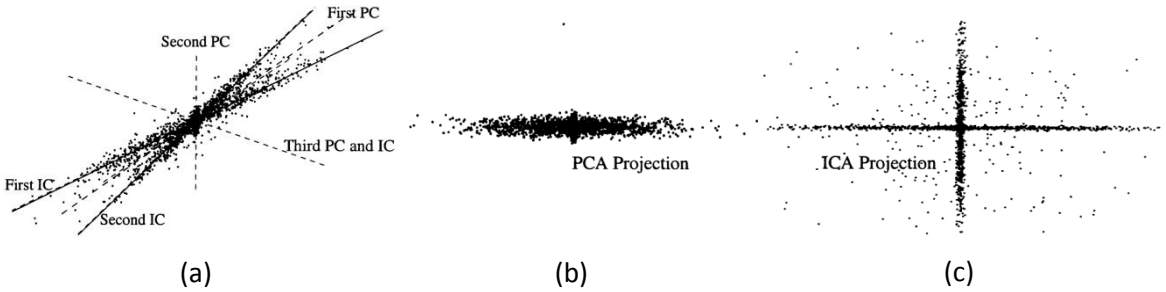

FIGURE 3.3: Distribution de données : (a) les axes correspondants aux composantes principales (PC) et aux composantes indépendantes (IC) (b) le sous-espace avec deux composantes basé sur l'ACP (c) idem avec l'ICA (source : [12]).

Toutes les méthodes de projection proposées précédemment reposent sur la même hypothèse, qui considère que les propriétés intrinsèques des données visuelles complexes telles que les images de visages, représentées dans un espace de grande dimension, sont souvent de faible dimension. En se basant sur cette hypothèse, ces méthodes cherchent à extraire uniquement les quelques dimensions principales afin de permettre au système de reconnaissance de n'utiliser que les données pertinentes relatives à l'identité des personnes, tout en minimisant le coût de calcul.

Bien que ces méthodes holistiques aient eu beaucoup de succès, leur inconvénient est qu'elles utilisent l'apparence globale du visage. Or, une telle représentation est sensible aux changements d'illumination et de pose. Une manière d'éviter ce problème consiste à utiliser des représentations locales du visage. En effet, les caractéristiques locales ne sont généralement pas aussi sensibles aux changements d'apparence que les caractéristiques globales.

3.2.2 Méthodes locales

Les méthodes locales peuvent être classées en deux catégories selon qu'elles sont basées sur l'extraction des composants caractéristiques du visage (par exemple : yeux, nez, etc.) ou non. Le premier travail de recherche en reconnaissance faciale, développé par Bledsoe dans les années 1960 [24], consiste en une méthode locale basée sur l'extraction manuelle de 20 points caractéristiques, ainsi que des mesures associées, à partir d'une photographie. Par la suite, beaucoup d'autres méthodes basées sur le même principe ont été proposées. Dans les méthodes de cette catégorie, la première étape consiste à extraire les composants caractéristiques du visage (comme les yeux, le nez, etc.). La deuxième étape consiste à extraire des informations géométriques ou d'apparence au niveau de chaque région. Une automatisation de l'extraction manuelle utilisée dans les travaux de Bledsoe a été proposée par Kanade [75]. Celle-ci met en œuvre un système d'extraction des points caractéristiques basés sur la détection de contours. Plus tard, Brunelli *et al.* ont proposé d'utiliser 35 paramètres géométriques pour modéliser un visage [29] (voir la Figure 3.4). La distance de Mahalanobis a été utilisée dans cette approche afin de mettre en correspondance deux visages.

Bien que ces méthodes soient simples et que leur coût de stockage soit très réduit, l'uti-

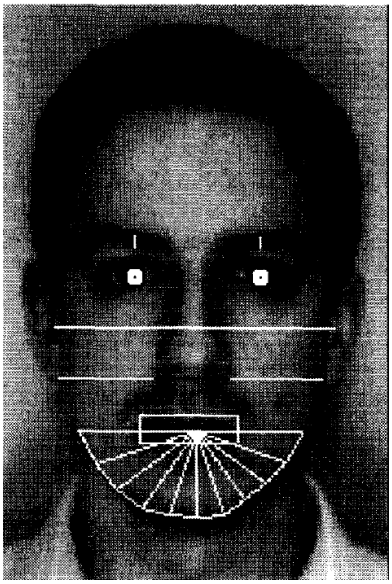

FIGURE 3.4: Mesures géométriques utilisées par Brunelli *et al.* [29].

lisation de quelques paramètres géométriques n'est pas suffisante pour représenter les variations des visages. Takacs a proposé d'utiliser des contours pour l'identification de visages [129]. Un filtre de Sobel est utilisé afin d'obtenir des cartes de contours qui sont ensuite mises en correspondance afin d'identifier un visage donné. Par la suite, une extension de ce travail a été proposée par Gao *et al.* [52], dans laquelle les auteurs ont introduit un descripteur nommé *Line Edge Map (LEM)*, qui consiste en un ensemble de segments de droites regroupant les pixels des contours. Les auteurs ont montré que cette méthode est invariante aux changements d'éclairage. Cependant, les variations d'apparence induites par les changements d'expressions ou de pose influencent considérablement la précision des résultats. Les méthodes locales citées précédemment se basent uniquement sur des informations géométriques. Or, les valeurs des pixels, ignorées par ces méthodes géométriques, peuvent être très utiles afin de mieux distinguer les visages. Pour cela, des méthodes ont été proposées consistant à utiliser, en plus des éléments géométriques extraits de l'image, des informations d'apparence au niveau de chaque point caractéristique. La méthode la plus utilisée est connue sous le nom de *Elastic Bunch Graph Matching* (EBGM) [152]. Le visage est représenté par une grille topologique dont les nœuds représentent des points caractéristiques du visage. Chaque nœud contient un ensemble de 40 coefficients d'ondelettes de Gabor (5 échelles et 8 orientations), appelé *jet* (voir la Figure 3.5). Ainsi, la géométrie d'un objet est codée par les arêtes du graphe, alors que les nœuds codent les variations des niveaux de gris. Suite au succès de cette méthode, plusieurs autres approches basées sur ce principe ont ensuite été proposées [153, 102]. Bien que les méthodes locales basées sur les points caractéristiques, notamment celles basées sur les graphes, aient connu un succès considérable, l'étape centrale d'extraction des points caractéristiques sur laquelle ces méthodes se basent représente le principal inconvénient, et constitue en soit un domaine de recherche en plein essor [42].

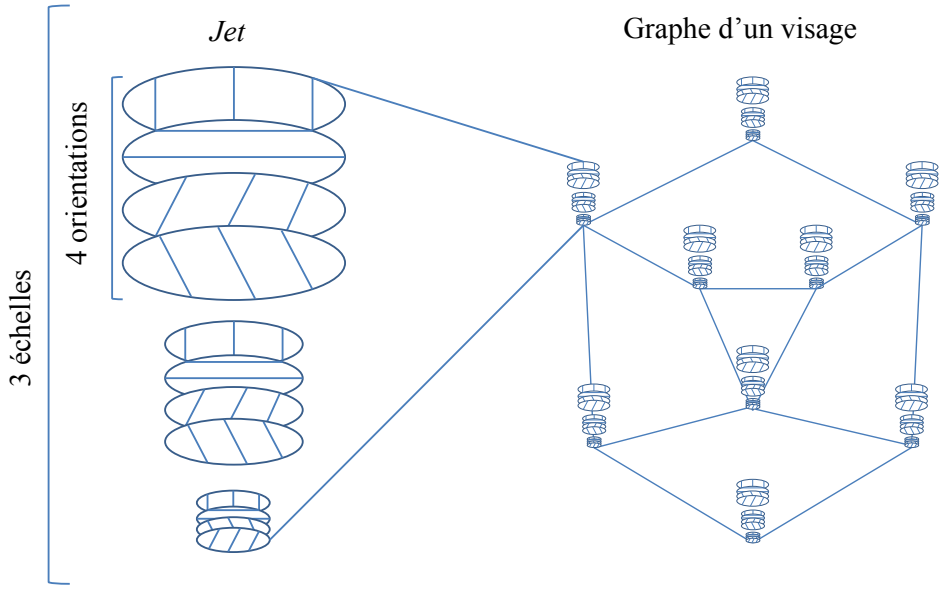

FIGURE 3.5: Exemple de représentation d'un visage par un graphe de 9 *jets* avec 4 orientations et 3 échelles (source : [152]).

La deuxième catégorie des méthodes locales consiste en un découpage du visage en différentes régions qui sont traitées localement. L'avantage de ces méthodes est qu'elles ne nécessitent pas une étape de détection des points caractéristiques du visage. Une méthode basée sur les modèles de Markov cachés (*Hidden Markov Model*) a été proposée par Samaria et Young [116]. Pour un visage frontal, une suite de régions ordonnées du haut vers le bas est extraite, et un état est associé à chaque région afin de construire un HMM. Les auteurs ont montré que leur méthode donne de bons taux de précision avec relativement peu de contraintes sur les données (comme les changements d'orientation, la non-homogénéité de l'éclairage, etc.). Une comparaison avec la méthode *Eigenfaces* a montré que leur méthode permet d'obtenir de meilleurs résultats. Une extension de cette méthode basée sur les modèles de Markov à deux dimensions a aussi été proposée par la suite afin de représenter les caractéristiques du visage horizontalement et verticalement [32]. Une autre méthode représentative de cette catégorie est l'utilisation des motifs binaires locaux (*Local Binary Pattern, LBP*) [4]. Le visage est divisé en un nombre de régions rectangulaires à partir desquelles les LBP sont extraits. Le principe de cette méthode est présenté en détails dans la Section 3.5.

Les méthodes locales ont prouvé leur efficacité pour la reconnaissance de visages, en raison notamment de leur robustesse à certains changements d'apparence. Cependant, cette robustesse reste limitée à un nombre restreint de facteurs de variations. Ainsi, les méthodes locales éparses comme EBGM sont robustes à certains changements d'expression, d'illumination et de pose, mais pas aux occultations ; en revanche, les méthodes probabilistes comme HMM sont robustes aux variations d'expression et aux occultations, mais pas aux changements de pose.

3.2.3 Méthodes hybrides

Les méthodes hybrides combinent des propriétés globales et locales, offrant ainsi potentiellement une meilleure représentation du visage. La méthode globale *Eigenfaces* [139] a été étendue à une version hybride par Pentland *et al.* [109]. Ces auteurs proposent d'appliquer le concept des *Eigenfaces* sur les régions faciales telles que les yeux et le nez afin d'obtenir des *Eigenfeatures*. Le visage est donc représenté globalement avec les *Eigenfaces* et localement avec les *Eigenfeatures*. Les auteurs ont montré que cette technique est plus précise que la technique strictement globale (*Eigenfaces*) ou celle strictement locale (*Eigenfeatures*). Le principe des *Fisherfaces* a aussi été étendu par Price et Gee [113] à une approche hybride qui consiste à utiliser une technique modulaire basée sur une variante de l'ADL. Le visage global est combiné à une bande faciale, de même largeur, s'étalant du front jusqu'au dessous du nez, et une autre bande faciale contenant uniquement les yeux. Les résultats montrent que cette approche est plus efficace que les techniques des *Eigenfaces* et des *Fisherfaces* en termes de robustesse aux changements d'éclairage du visage, d'expression faciale, et aux occultations partielles.

L'utilisation des modèles de visages [84] pour la reconnaissance de visages, comme les modèles actifs d'apparence (*active appearence model, AAM*) [38] constitue un autre exemple d'approche hybride. L'efficacité de ces modèles pour la reconnaissance de visages a été démontrée dans [46] et [83]. La reconnaissance de visages avec les AAM comprend deux étapes : la construction du modèle et l'ajustement de celui-ci sur de nouveaux visages.

- **Construction** : la modélisation du visage est une étape d'apprentissage qui consiste à générer un modèle statistique représentant les variations de la forme et de la texture d'un ensemble d'images de visages constituant la base d'apprentissage. Les images sont toutes annotées manuellement par un nombre donné de points caractéristiques marqués sur chaque visage. Afin de générer le modèle de forme, les ensembles de points caractéristiques marqués sur chaque visage sont alignés dans un système de coordonnées commun et sont représentés par des vecteurs. Une ACP est ensuite appliquée sur l'ensemble de ces vecteurs. La construction du modèle de variation d'apparence est effectuée en déformant d'abord la forme de chaque visage afin qu'elle soit alignée à la forme moyenne. Le vecteur d'apparence est ensuite extrait et une ACP est appliquée afin d'obtenir un modèle linéaire. Les deux modèles peuvent être utilisés séparément [83] ou conjointement [46].
- **Ajustement** : une fois que le modèle est construit, la deuxième étape consiste à ajuster les paramètres du modèle selon une image donnée en cherchant l'ensemble optimal des paramètres qui représentent le mieux l'image en entrée. Ceci est considéré comme un problème d'optimisation où l'on essaie de faire varier les paramètres du modèle tout en minimisant la distance entre le modèle et l'image. Pour cela, une étape d'apprentissage est d'abord appliquée afin de permettre au modèle d'apprendre comment résoudre le problème d'optimisation a priori. Elle consiste en un apprentissage à partir d'un ensemble d'images annotées où les paramètres du modèle sont connus. Pour chaque exemple dans l'ensemble d'apprentissage, un certain nombre de déplacements connus sont appliqués au modèle, et le vecteur de différences est enregistré. Quand les

données d'apprentissage sont suffisantes, une régression multiple est appliquée pour modéliser la relation entre les déplacements du modèle et les vecteurs de différences. L'AAM encode donc des informations sur la manière dont les paramètres doivent être ajustés pour atteindre le meilleur alignement. Étant donnée une image d'un visage à reconnaître, le modèle est d'abord placé sur le visage et le calcul du vecteur de différences est réalisé. Ensuite, l'ajustement est effectué en utilisant le modèle de régression obtenu dans la phase d'apprentissage : celui-ci prédit le mouvement qui donne la meilleure correspondance. Le processus est répété jusqu'à la convergence. La Figure 3.6 montre deux exemples d'ajustement d'un AAM sur un visage après différents nombres d'itérations.

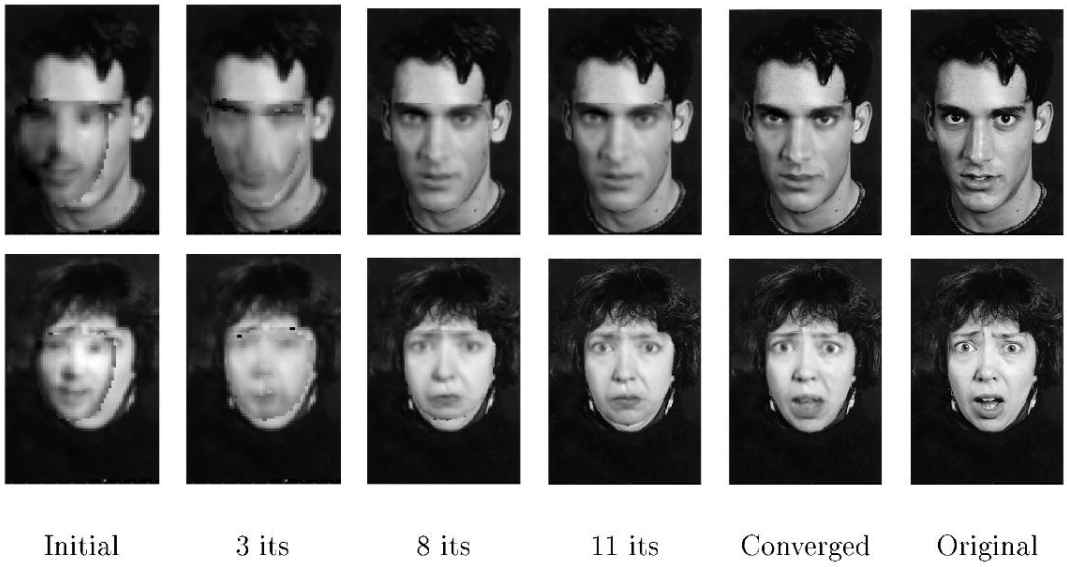

FIGURE 3.6: Deux exemples d'ajustement d'un AAM à une image de test (source : [38]).

L'algorithme de reconnaissance de visages basé sur les AAM est présenté en détails dans [40]. L'efficacité de ces méthodes est fortement liée à la richesse de la base d'apprentissage utilisée pour construire le modèle. Ce dernier devrait être aussi complet que possible pour être capable de synthétiser une approximation suffisamment proche de toute image du visage de la personne cible.

3.2.4 Synthèse

Durant ces dernières années, plusieurs méthodes ont été proposées pour la reconnaissance de visages à partir d'images 2D. Parmi les approches globales, les *Eigenfaces* [139] et les *Fisherfaces* [13] ont prouvé leur efficacité et ont été beaucoup utilisées dans l'état de l'art [163]. Les méthodes basées sur des caractéristiques locales [152] ont également eu beaucoup de succès. Par rapport aux approches globales, elles sont moins sensibles aux variations d'éclairage et de pose. Cependant, les techniques d'extraction de caractéristiques nécessaires pour ce type d'approche ne sont pas encore assez fiables ni assez précises. Par conséquent,

les deux types d'informations globales et locales sont importants pour la reconnaissance de visages.

Dans toutes les approches présentées jusqu'ici, les images de visages sont acquises à l'aide de caméras, qui fournissent des images bidimensionnelles qui les caractérisent. Cette réduction de la dimension de l'information 3D du visage fait ressortir quelques problèmes qui se posent lorsque les images 2D ne sont pas parfaitement frontales où lorsque les conditions environnementales d'acquisition ne sont pas contrôlées. Les changements de pose et d'éclairage sont les deux problèmes principaux engendrés par cette projection. Plusieurs méthodes ont été proposées pour minimiser la sensibilité des méthodes de reconnaissance 2D aux différentes variations du visages comme les changements d'expressions [132], d'éclairage [93] ou de poses [39, 109]. Cependant, l'invariance à ces changements reste toujours un problème non résolu. En effet, la perte considérable d'informations 3D du visage constitue une limitation majeure des méthodes de reconnaissance 2D. Afin de résoudre ce problème, de nombreuses méthodes basées sur la forme 3D du visage ont été proposées. Dans la section suivante, nous nous intéressons à la reconnaissance 3D du visage.

3.3 Approches de reconnaissance 3D

Après plus de 40 années de recherche et de développements, la reconnaissance de visages 2D a atteint un certain niveau de maturité dans des environnements contrôlés. Toutefois, lorsqu'il s'agit d'un contexte non contrôlé où des variations (environnementales ou de visage) sont susceptibles de se produire, cette tâche reste encore très difficile à effectuer. Par exemple, les évaluations de FERET et FRVT ont révélé qu'il persiste au moins deux défis majeurs : le problème des variations d'éclairage et de la pose. La reconnaissance 3D du visage permet de donner une solution aux problèmes rencontrés par les approches de la reconnaissance 2D classiques. Ces techniques sont basées sur le modèle 3D du visage où la forme est totalement conservée, à la différence de la reconnaissance 2D où une perte d'informations considérable est engendrée par la projection. Les méthodes de reconnaissance 3D peuvent être classées en quatre catégories : les méthodes d'alignement, les méthodes basées sur des caractéristiques géométriques, les méthodes de réduction de dimensionnalité d'espace et enfin les méthodes basées sur des modèles 3D du visage.

3.3.1 Méthodes d'alignement

Ces méthodes résolvent le problème de la reconnaissance 3D de visages en alignant les surfaces 3D qui modélisent les deux visages à comparer (voir la Figure 3.7). L'algorithme généralement utilisé est l'algorithme du plus proche point itératif (*iterative closest point, ICP*), qui a été introduit par Besl *et al.* [16]. Il consiste en une optimisation alternée d'appariements et de transformations. Ainsi, à partir d'une transformation initiale, les deux étapes suivantes sont itérées :

– mise en correspondance : on compare chaque primitive du modèle transformé avec la primitive la plus proche dans la référence ;

FIGURE 3.7: Exemple d'alignement de deux visages 3D.

– recalage : la transformation (translation + rotation) est calculée pour minimiser l'erreur quadratique.

Colbry et Stockman [37] ont proposé un système appelé 3DID qui utilise l'erreur quadratique moyenne (*root mean squared error*) obtenue par un alignement par ICP pour mesurer la différence entre les modèles 3D. Ils ont montré qu'il existe un seuil au-delà duquel cette erreur indique que le visage appartient à la même personne. Un inconvénient de l'ICP indiqué dans ce travail est le temps de calcul très élevé nécessaire pour comparer un visage de test avec tous les visages enregistrés. Une solution à ce problème a été proposée par Bardsley *et al.* [10]. Elle consiste à construire un modèle moyen de tous les visages de la collection. Un visage de test est aligné uniquement avec ce modèle et l'erreur moyenne du point au plan (*average point-to-plane error*) est utilisée comme métrique de reconnaissance. Dans les méthodes utilisant l'ICP, le visage est considéré comme un objet rigide. Ainsi, les changements d'expression faciale influence beaucoup la performance de ces méthodes. Dans [6], les auteurs proposent d'introduire une nouvelle métrique orientée région dans l'algorithme de l'ICP. Il s'agit de segmenter un visage en plusieurs régions ayant des influences plus ou moins importantes sur la déformation de la forme 3D du visage. Le calcul de la similarité est pondéré en fonction de ces régions en donnant plus de poids aux régions statiques qu'aux régions dynamiques du visage.

3.3.2 Méthodes basées sur des propriétés géométriques

Afin d'extraire des informations plus représentatives de la forme 3D du visage, plusieurs méthodes basées sur les propriétés géométriques 3D ont été définies. Un des premiers travaux dans cette catégorie a été proposé par Lee et Milios [87], dans lequel les régions convexes du visage sont sélectionnées et représentées par une image gaussienne étendue (*extended gaussian image, EGI*). L'image gaussienne étendue représente l'information de la normale à la surface, projetée dans un espace sphérique. La similarité entre deux visages est évaluée par la corrélation des images EGI les représentant. Ce choix est basé sur le fait que les régions convexes sont moins sensibles aux variations des expressions faciales. L'utilisation des EGI a été étendue par Tanaka *et al.* [131]. Dans leur approche, les auteurs calculent des cartes de courbures principales minimales et maximales à partir des cartes de profondeur des visages.

Ensuite, ils extraient les lignes de sommets et celles de vallées sur la surface faciale, qui sont respectivement des vecteurs correspondant aux maxima locaux sur les valeurs de courbure principale minimale et aux minima locaux sur les valeurs de courbure principale maximale. Ensuite, des images gaussiennes étendues sont construites en mettant en correspondance les vecteurs de chaque type (sommets et vallées) sur une sphère unité. La similarité entre deux cartes de profondeur est évaluée en faisant la moyenne de la corrélation des EGI des sommets et des vallées. Chua *et al.* [36] utilisent les signatures de points (*points signatures*) pour localiser les points de référence sur un visage 3D. Ces signatures sont invariantes aux changements de pose, et sont donc utilisées pour normaliser cette dernière. Afin de faire face au problème de changements d'expressions, seules les régions rigides (parties allant du dessous du nez jusqu'au front) sont utilisées pour la mise en correspondance. Dans [103], Moreno *et al.* proposent de subdiviser le visage en 86 régions en utilisant un algorithme de segmentation qui exploite la médiane et la courbure gaussienne pour éliminer les régions ayant une courbure importante. Puis un vecteur caractéristique pour chaque visage est créé à partir des régions non éliminées.

Par rapport aux approches géométriques présentées précédemment, les techniques 3D d'identification du visage basées sur les points caractéristiques locaux restent relativement peu développées. Dans [89], Lee *et al.* proposent un système de reconnaissance 3D de visages où, à partir des mesures 3D, huit points caractéristiques du visage géométriquement invariables sont extraits puis utilisés pour calculer un vecteur caractéristique décrivant les distances entre ces points et les angles qu'ils forment.

Une méthode a été proposée par Bronstein *et al.* [28] où le visage est traité comme un objet déformable dans la géométrie de Riemann muni d'une distance géodésique, qui est beaucoup moins sensible aux changements d'expressions faciales que la distance euclidienne. L'idée consiste à représenter la surface faciale par une surface isométrique appelée forme canonique (*canonical form*). Les points 3D du visage (munis initialement d'une distance géodésique) sont projetés vers un espace de dimension réduite muni d'une distance euclidienne, tout en préservant la distance géodésique. Ces formes canoniques sont par la suite alignées et interprétées sur une grille cartésienne, donnant naissance à des images canoniques. Finalement, un espace propre est créé à partir de ces images canoniques d'apprentissage. Les images de test sont assujetties aux mêmes traitements avant d'être mises en correspondance.

Les propriétés géométriques de la surface 3D utilisées par les méthodes de cette catégorie offrent une bonne discrimination de l'information pour des finalités de reconnaissance de visage. Cependant, un inconvénient majeur de ces méthodes est qu'elles sont très sensibles au bruit du fait qu'elles sont souvent basées sur des calculs de courbures (dérivées secondes). Ceci entraîne des erreurs de localisation des caractéristiques qui peuvent être aggravées par des occultations partielles résultant d'un changement de pose.

3.3.3 Méthodes de réduction de dimensionnalité

Comme dans la reconnaissance 2D, les méthodes statistiques de réduction d'espace ont également été appliquées sur les données 3D. Une étude comparative a été proposée par Chang *et al.* [31] où deux scénarios de reconnaissance basée sur l'ACP ont été suivis : un

premier en utilisant des images d'intensité et un deuxième en utilisant des images de profondeur. Ils ont rapporté une meilleure précision dans le cas de l'utilisation des cartes de profondeur. De même, Hesher *et al.* [61] ont proposé un système de reconnaissance basé sur l'application de l'ACP sur des données de profondeur. Les auteurs ont montré que la précision obtenue n'est pas satisfaisant notamment quand les images de profondeur sont bruitées. Heseltine *et al.* [60] ont développé une approche qui applique l'ACP sur des représentations tridimensionnelles du visage. Elle consiste à calculer des surfaces propres (*Eigensurface*) à partir de modèles 3D maillés de visage. Une extension de la méthode *Ficherfaces* [13] dite *Fishersuface* a été aussi appliquée aux données surfaciques de visages 3D. L'ACP a été aussi combinée avec les modèles de Markov cachés [2]. D'autres approches basées sur l'Analyse Discriminante Linéaire ou l'Analyse des Composantes Indépendantes ont aussi été développées pour l'analyse des données 3D de visages [54, 79].

Les méthodes de réduction de dimensionnalité (comme l'ACP) appliquées aux données 3D, montrent des résultats moins convaincants que leur application aux images 2D ; elles sont souvent utilisées en tant que méthodes de référence lors des études comparatives [110].

3.3.4 Méthodes basées sur les modèles 3D

Le principe de ces méthodes consiste à utiliser un modèle 3D paramétrique qui représente les différentes variations du visage. Le modèle le plus utilisé dans l'état de l'art est le modèle déformable 3D (*3D Morphable Model, 3DMM*) [22]. Il s'agit d'un modèle paramétrique basé sur une représentation du visage dans un espace vectoriel obtenu à partir d'un ensemble de numérisations 3D. Le 3DMM permet de synthétiser une nouvelle image 2D du visage, qui ressemble à l'image en entrée. Le modèle déformable peut aussi être utilisé de différentes manières pour la reconnaissance de visages. Dans [23], étant donnée une image du visage, Blanz *et al.* proposent un algorithme d'optimisation qui estime les paramètres de forme et d'apparence du 3DMM donnant un rendu proche de cette image. Ces paramètres sont utilisés pour représenter un visage donné. L'identification d'un visage est obtenue en calculant la distance entre les paramètres estimés pour ce visage et l'ensemble des paramètres enregistrés dans la base des visages connus. La Figure 3.8 illustre le fonctionnement de cette méthode. Le 3DMM peut aussi être utilisé pour générer un nombre important d'images synthétiques sous différents points de vue. Partant de l'image 2D de test, un processus de déformation du modèle est appliqué afin de produire une approximation qui représente cette image. Ensuite, un ensemble de rotations est appliqué afin de générer les différentes vues du visage pour chaque sujet. Dans [150], des images avec différentes poses sont générées à partir d'un 3DMM en vue de produire un corpus d'apprentissage plus important englobant différentes variations de pose. Dans [142], Wang *et al.* utilisent l'ensemble des images générées afin de calculer la distance minimale entre l'image de test et les différents ensembles d'images pour chaque personne. Une autre méthode consiste à utiliser le 3DMM pour obtenir une vue frontale à partir d'une image 2D acquise sous différents angles de vue. L'approche basée sur le 3DMM a été utilisée dans l'évaluation FRVT et sa performance a été démontrée dans le cas de la reconnaissance de visages sous différentes variations de poses [111]. Dans [108], une version différente du 3DMM appelée *Basel Face Model* (BFM) a été proposée. Les au-

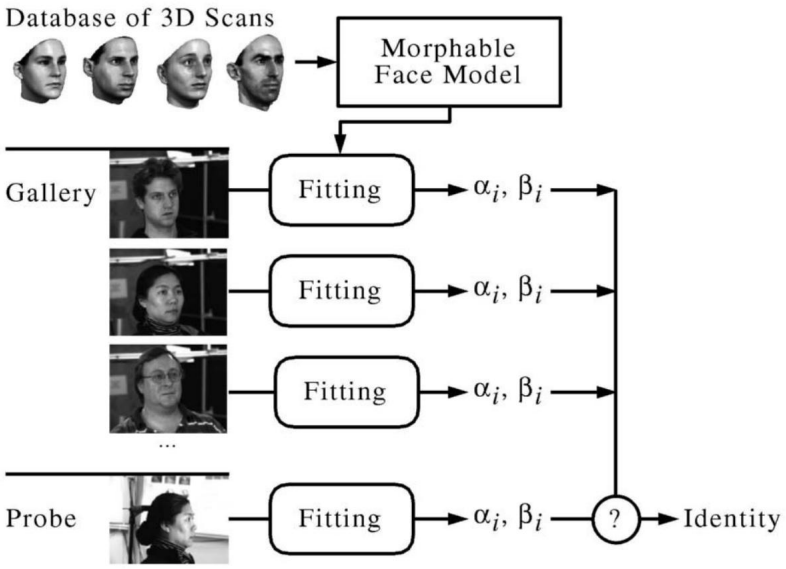

FIGURE 3.8: Reconnaissance de visages à l'aide du 3DMM (source : [23]).

teurs ont démontré l'invariance de ce nouveau modèle aux différents types de changements de pose et d'illumination. Le BFM est disponible publiquement et une plateforme a été mise en place permettant l'évaluation de différentes méthodes sur les mêmes données.

Les bons résultats des approches présentées ici s'expliquent par le fait qu'un modèle 3D est plus complet et plus puissant qu'une image 2D pour prendre en compte les conditions d'éclairage et de pose. Cependant, le temps de traitement des méthodes basées sur le 3DMM est très élevé et une haute qualité d'images est nécessaire afin de pouvoir déformer le modèle. La construction d'un modèle 3D déformable est complexe. Elle nécessite de collecter un nombre important de scans 3D de visage ainsi que de les annoter manuellement [165]. Un autre inconvénient de ces méthodes est que les visages reconstruits sont susceptibles d'être plus proches du modèle moyen que de leur propre modèle si la base utilisée pour la reconstruction du modèle est petite ou contient peu de variations (pose, expression, etc.).

3.3.5 Synthèse

La reconnaissance 3D, bien qu'elle soit beaucoup plus précise et surtout invariante aux changements d'éclairage en comparaison avec la reconnaissance 2D, souffre de quelques problèmes. La sensibilité aux changements d'expressions faciales est un inconvénient majeur des méthodes 3D. Des méthodes basées sur des régions convexes et rigides ont été proposées pour le résoudre, mais elles sont généralement basées sur des calculs géométriques très sensibles au bruit et au paramétrage de la segmentation. L'utilisation de la forme 3D du visage ne résout pas le problème des occultations, et par conséquent différentes solutions sont aussi proposées pour contourner ce problème [44]. Bien que les méthodes 3D aient permis de lever quelques verrous scientifiques rencontrés en reconnaissance 2D, elles ignorent

totalement l'information d'apparence visuelle. Par ailleurs, l'efficacité des méthodes 3D est fortement liée à la qualité des données 3D utilisées. Dans l'état actuel des recherches, aucun consensus ne se dégage quant à la modalité (2D ou 3D) qui donne les meilleurs résultats. En effet, la précision de reconnaissance qu'il est possible d'atteindre avec une modalité est conditionnée par le contexte d'utilisation, la nature et la qualité des données, et la méthode de reconnaissance utilisée. L'utilisation conjointe des données 2D et 3D s'avère être une piste pertinente pour tirer parti de ces deux modalités complémentaires.

3.4 Approches bimodales 2D-3D

Afin d'augmenter la précision et la robustesse des systèmes de reconnaissance faciale, de nouvelles méthodes bimodales 2D-3D, dites de fusion, ont été récemment développées [26, 1]. L'objectif de ces méthodes est de combiner l'information visuelle (image 2D) et l'information 3D correspondante (modèle 3D ou image de profondeur du visage) afin de tirer parti des avantages et de la complémentarité des deux modalités. L'image 2D fournit des informations sur les régions texturées du visage avec peu de structure géométrique (comme les poils du visage, les yeux et les sourcils), tandis que les données 3D fournissent des informations sur les régions où il y a peu de texture (comme le nez, le menton ou les joues). La fusion de ces deux modalités est donc susceptible d'améliorer la précision et la robustesse des méthodes de reconnaissance faciale. La fusion 2D-3D peut intervenir à différents niveaux du processus de reconnaissance. Trois stratégies de fusion peuvent être considérées selon le niveau auquel elles interviennent lors du processus de reconnaissance [59] :

- **la fusion de données brutes** : elle consiste à combiner les données provenant directement des capteurs afin de construire de nouvelles données.
- **la fusion de descripteurs** : elle consiste à modéliser les données de chaque modalité séparément. Les vecteurs caractéristiques extraits à partir des données de chaque modalité sont ensuite fusionnés afin d'en construire un seul qui va être utilisé lors de l'apprentissage et de la mise en correspondance.
- **la fusion de décisions** : elle intervient après l'étape de classification. Un classifieur par modalité est donc construit et leurs sorties respectives sont ensuite combinées.

Les Figures 3.9, 3.10 et 3.11 illustrent les trois stratégies possibles pour la fusion lors de l'étape de test. Les méthodes de fusion peuvent aussi être classées selon qu'elles interviennent avant ou après l'étape de classification [117]. On parle donc de fusion *précoce* (fusion de données ou de descripteurs) ou *tardive* (fusion de décisions).

3.4.1 Fusion de données brutes

Cette stratégie intervient au niveau des données brutes obtenues par les différents capteurs. La fusion des données brutes a été peu utilisée dans la reconnaissance 2D-3D. Papatheodorou *et al.* [106] ont proposé une généralisation de l'algorithme d'alignement ICP en 4D (x, y, z, intensité). Les expérimentations effectuées ont montré une précision de reconnaissance élevée dans le cas d'une vue frontale avec une expression neutre du visage. Dans

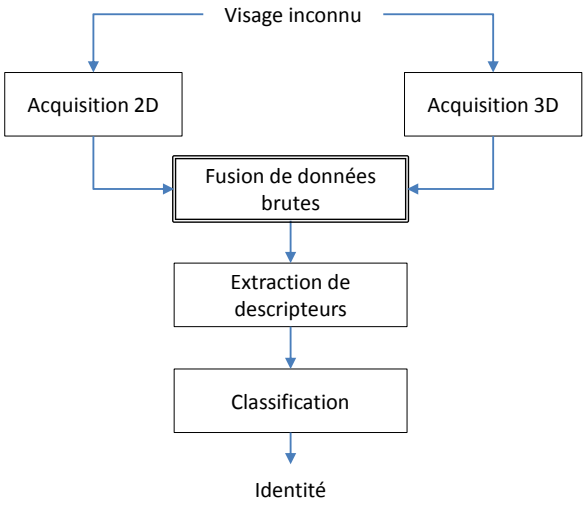

FIGURE 3.9: Fusion de données brutes (précoce, avant l'extraction de descripteurs).

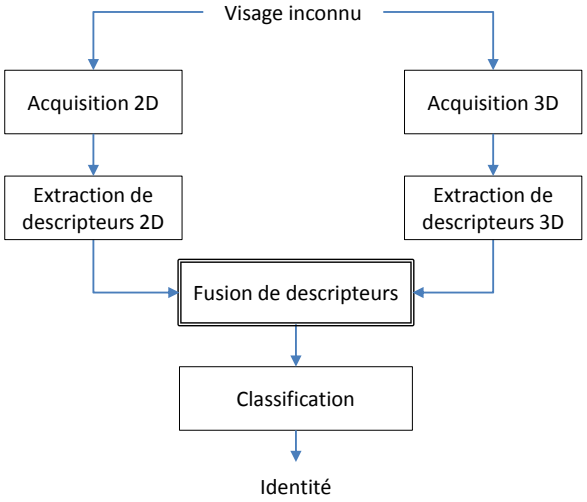

FIGURE 3.10: Fusion de descripteurs (précoce, après l'extraction de descripteurs).

[82], les données 2D et 3D sont représentées par les images d'intensité et de profondeur respectivement. Les deux images sont fusionnées en appliquant une ACP pour déterminer les axes orthogonaux qui décorrèlent les données. Le nouvel espace créé par l'ACP permet de fusionner linéairement les données 2D et 3D en deux ensembles de données décorrélées. Hajati *et al.* [58] appliquent une transformation sur les pixels des images 2D selon les distances géodésiques correspondantes dans les images de profondeur.

L'avantage de ce mode de fusion est qu'il permet d'exploiter la dépendance entre les deux modalités. De plus, au niveau des données, on dispose d'informations très riches et complémentaires sur le visage, dont l'étape d'extraction de descripteurs peut pleinement bénéficier. Cependant, les données brutes sont souvent bruitées, et une fusion à ce niveau

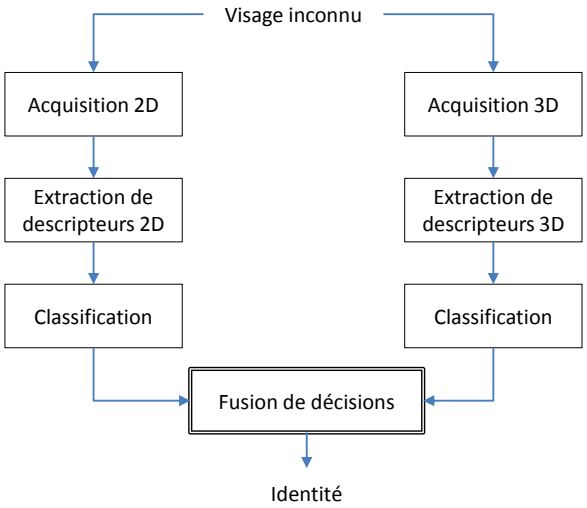

FIGURE 3.11: Fusion de décisions (tardive).

signifie que les descripteurs extraits de données mélangeant deux types de bruit provenant des deux modalités. De plus, les données 2D et 3D sont de natures différentes. Il est donc plus adéquat de calculer des descripteurs spécifiques pour chaque modalité.

3.4.2 Fusion de descripteurs

La fusion de descripteurs a été plus utilisée dans les méthodes bimodales de reconnaissance de visages que la fusion de données brutes. Cette stratégie de fusion consiste en deux étapes. Dans la première étape, les vecteurs caractéristiques sont calculés pour chaque modalité, soit en utilisant le même type de descripteur [156], soit en utilisant des descripteurs de types différents [7, 99, 146]. Dans le deuxième cas, Le nombre et la dynamique des composants dans les deux descripteurs peuvent être différents. Un descripteur avec un grand nombre de composants ou ayant des composants avec une dynamique forte peut écraser l'autre. Une étape de normalisation est nécessairement appliquée avant la fusion des descripteurs afin de ramener les valeurs des deux descripteurs à une même échelle. Des méthodes de réduction d'espace, notamment l'ACP, sont aussi souvent appliquées sur les descripteurs avant leur fusion [146] afin de réduire la dimension du vecteur caractéristique final. La deuxième étape consiste à combiner les descripteurs des deux modalités afin d'en construire un seul. Ceci peut être effectué par une simple concaténation des deux vecteurs [146, 99]. Quelques approches appliquent des méthodes de sélection de caractéristiques afin de fusionner les descripteurs [91, 156]. Li *et al.* [91] proposent une méthode d'apprentissage basée sur AdaBoost pour combiner des histogrammes LBP calculés sur les données 2D et 3D. Ces histogrammes sont concaténés en un seul vecteur, et ensuite AdaBoost est appliqué afin de sélectionner les caractéristiques pertinentes. Dans les travaux de Xu *et al.* [156], une ADL est d'abord effectuée afin de réduire la dimension des vecteurs caractéristiques 2D et 3D et ensuite AdaBoost est appliqué pour combiner les caractéristiques 2D et 3D les plus pertinentes.

L'avantage de la fusion de descripteurs par rapport à celle des données est que le bruit est réduit par le calcul des descripteurs. L'utilisation de descripteurs appropriés pour chaque modalité permet une meilleure représentation de celles-ci. Cependant, un problème central de ce mode de fusion est la grande dimensionnalité du descripteur fusionné qui peut conduire au problème de la malédiction de la dimension (plus de variables que d'observation) et qui peut aussi entraîner un coût de calcul élevé.

3.4.3 Fusion de décisions

Le fusion de décisions est la stratégie la plus utilisée dans la reconnaissance bimodale 2D-3D de visages [1, 26]. Elle intervient après l'étape de mise en correspondance des descripteurs. Elle peut être effectuée au niveau des scores (i.e. mesures de similarité par rapport à chaque classe) [31, 126, 138] ou de décisions finales (i.e. l'identité) [53] obtenues via les différents systèmes mono-modaux. Concernant la fusion de scores, il est important d'effectuer au préalable une étape de normalisation afin de projeter les scores obtenus par les différents classifieurs dans le même intervalle [123]. La fusion des scores peut être effectuée selon différentes règles comme la somme, le produit, la somme pondérée, etc. Dans [31], la somme a été utilisée pour combiner deux distances de Manahalobis calculées dans deux espaces créés par une ACP pour les données 2D et 3D séparément. Le produit des scores a été utilisé pour combiner des distances euclidiennes [138] et des distances de Manahalobis obtenues par la méthode *Eigenfaces* [126, 138]. Par ailleurs, la technique la plus utilisée pour la fusion des scores est la somme pondérée [14, 17, 71, 70, 123, 137, 141]. Elle consiste à attribuer des poids aux classifieurs utilisés pour la fusion. Ces poids sont déterminés par différentes techniques. Dans [137], les poids sont choisis expérimentalement lors de la phase d'apprentissage des classifieurs. Dans [17], la méthode de Fisher a été utilisée afin de combiner linéairement les mesures de similarité obtenues par quatre classifieurs (deux pour les données 2D et deux pour les données 3D). Cette même technique a été utilisée pour définir les poids des classifieurs dans [14].

Concernant la fusion de décisions finales, elle intervient plus tard dans le processus de reconnaissance. Chaque classifieur fournit une seule décision quant à la classe représentant l'identité du visage. La fusion de décisions finales est souvent obtenue par une méthode de vote majoritaire ou encore de vote majoritaire pondéré [53], où des coefficients de confiance sont attribués aux classifieurs selon leur précision pour une décision donnée.

La fusion de décisions est plus simple du fait que les deux modalités sont traitées d'une manière indépendante, et que seules les sorties sont considérées lors de la fusion. Cependant, les systèmes basés sur cette approche peuvent donner la même qualité de résultats que celle obtenue via une seule modalité. En effet, un des classifieurs peut dominer les autres, et être seul à l'origine de la précision globale du système. La plupart des méthodes basées sur ce type de fusion considèrent que les deux modalités sont indépendantes. Cependant, l'hypothèse de l'indépendance des représentations 2D et 3D est discutable car les données sont extraites du même visage. Comme il a été mentionné par Husken *et al.* [70], la position des traits du visage (yeux, nez, etc.) est la même. La fusion de décisions ne permet pas d'opérer de manière synergique du fait que chaque modalité est traitée à part.

3.4.4 Choix de la stratégie de fusion

Chacune des stratégies a ses avantages et ses inconvénients, et il est difficile de déterminer la stratégie la plus adéquate à utiliser pour la fusion des deux modalités. La précision d'une stratégie peut être liée à la méthode de reconnaissance utilisée. Par exemple, dans [14], Benabdelkader *et al.* ont proposé deux approches de reconnaissance avec deux stratégies de fusion différentes. Dans la première approche, le visage est représenté par un graphe reliant ses points caractéristiques. La concaténation des données ou des descripteurs peut donc perturber la structure du visage. Par conséquent, c'est la fusion de décisions qui a été choisie. La deuxième approche est basée sur les *Fisherfaces*. La fusion de descripteurs est choisie par les auteurs en se basant sur l'hypothèse que les données 2D et 3D sont corrélées (elles ne sont pas orthogonales), et que la fusion de descripteurs permet d'en tenir compte, ce qui n'est pas le cas pour la fusion de décisions. Dans [123], Soltana *et al.* déterminent un lien entre le niveau de fusion et la relation entre les modalités. Ils considèrent que la fusion des données ou des descripteurs est une fusion complémentaire et celle de décisions est compétitive. Il est donc important de comprendre la relation entre les données 2D et 3D du visage afin de choisir la meilleure stratégie de fusion. Dans une étude théorique proposée par Durrant-Whyte [45] sur la fusion multi-capteurs, l'auteur définit des capteurs comme complémentaires s'ils ne sont pas directement dépendants mais peuvent être combinés pour avoir une image complète du phénomène observé. Par ailleurs, dans cette étude, les capteurs sont considérés compétitifs s'ils fournissent des mesures indépendantes pour la même propriété. Selon cette définition, il est difficile de décider si les capteurs 2D et 3D sont complémentaires ou compétitifs. Afin d'éclaircir cette ambiguïté, des études expérimentales sont donc menées afin de comparer les différentes stratégies [14, 91]. Cependant, les résultats obtenus par ces études ne sont pas concluants. En effet, la fusion de décisions permet d'atteindre de meilleurs résultats que la fusion de descripteurs dans les travaux de Benabdelkader *et al.* [14], alors que le contraire a été rapporté dans les travaux de Li *et al.* [91].

3.4.5 Synthèse

Les méthodes bimodales de reconnaissance de visages ont toutes montré l'intérêt de la fusion des données 2D et 3D afin d'obtenir un système plus précis et plus robuste. Il ressort des études présentées ici que lorsque ces données sont fusionnées selon différentes stratégies, un taux de reconnaissance supérieur aux méthodes monomodales (2D ou 3D) peut être atteint. Par ailleurs, même si le choix de la meilleure stratégie de fusion n'est pas toujours évident pour la plupart des approches bimodales 2D-3D, ce choix est crucial afin de pouvoir bénéficier au mieux de la complémentarité des deux modalités, et ainsi améliorer les résultats [110].

3.5 Reconnaissance 3D basée sur les motifs binaires locaux

Comme la forme 3D du visage peut être représentée par une image de profondeur, plusieurs travaux ont été menés consistant à appliquer des descripteurs 2D utilisés initialement

dans la reconnaissance 2D, comme les filtres de Gabor, les descripteurs de Haar ou les motifs binaires locaux (*Local Binary Pattern, LBP*), sur les images de profondeur ont été proposées [91, 148, 156]. En raison de sa simplicité et son efficacité pour la reconnaissance 2D du visage, le descripteur LBP a été utilisé dans de nombreux travaux récents de reconnaissance 3D. Dans ce travail de thèse, nous nous intéressons à ce descripteur calculé sur des images de profondeur pour la représentation de la forme 3D du visage. Nous présentons ainsi dans cette section son principe et les méthodes de reconnaissance 3D de visages basées sur ce descripteur.

3.5.1 Principe

Le descripteur LBP a été proposé initialement par Ojala *et al.* [104, 105]. L'opérateur décrit chaque pixel par la valeur relative des niveaux de gris des 8 pixels voisins (voir la Figure 3.12). Si la valeur du niveau de gris du pixel voisin est supérieure ou égale à celle du pixel central, on lui attribue la valeur de 1, sinon 0. Les valeurs binaires associées aux voisins sont alors lues de façon séquentielle, dans le sens horaire depuis un point de référence, pour former une suite binaire qui est utilisée pour caractériser la texture locale.

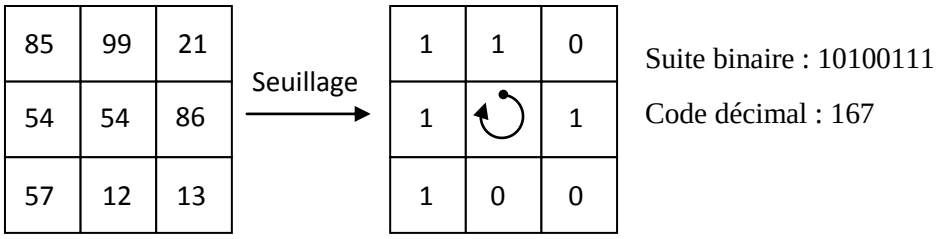

FIGURE 3.12: Exemple de calcul du code LBP original pour un pixel d'une image (calcul pour le pixel central).

Le descripteur LBP original a été étendu par la suite à différentes tailles de voisinage et différents rayons pour intégrer des changements d'échelle [104]. Le voisinage local d'un pixel est défini comme un ensemble de points régulièrement espacés sur un cercle centré sur le pixel donné (voir la Figure 3.13). On adopte la notation $LBP_{(R,V)}$ pour dénoter les LBP construits pour un rayon R et un voisinage de taille V. En faisant varier la valeur du rayon R, les LBP de différentes échelles sont ainsi obtenus. Le paramètre de voisinage V règle la résolution d'échantillonnage, et donc la taille du code final.

Formellement, le code $LBP_{(R,V)}$ est calculé de la manière suivante :

$$LBP_{(R,V)}(x,y) = \sum_{i=0}^{V-1} s(n_i - n_c)2^i, \; avec \; s(k) = \begin{cases} 1 & \text{si} \quad k \geq 0 \\ 0 & \text{sinon} \end{cases} \tag{3.1}$$

où :

– n_c correspond à la valeur du niveau de gris du pixel central du voisinage local ;

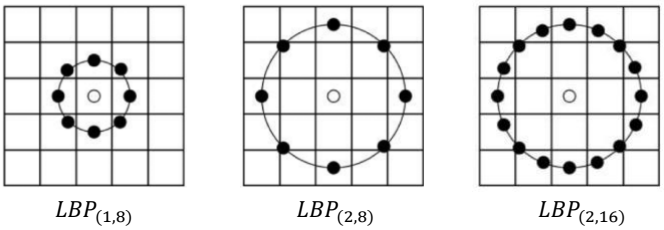

FIGURE 3.13: Exemples de voisinages $\mathrm{LBP}_{(R,V)}$ avec différentes valeurs de rayon R et de voisinage V.

– n_i correspond aux valeurs des niveaux de gris des pixels voisins situés autour du pixel central avec un rayon R. La position du pixel voisin est calculée par l'Equation 3.2. Dans le cas où cette position ne tombe pas exactement sur un pixel, la valeur du voisin est estimée par une interpolation bilinéaire, permettant ainsi différentes valeurs de rayons et différents nombres de points voisins.

$$(x_{p_i}, y_{p_i}) = \left(x_p + R\cos\left(\frac{i}{V} \times 2\pi\right), y_p - R\sin\left(\frac{i}{V} \times 2\pi\right)\right) \tag{3.2}$$

Dans la pratique, l'Equation 3.1 indique que les signes des différences dans un voisinage sont interprétés comme une valeur binaire de V bits, résultant en 2^V valeurs distinctes possibles pour le motif binaire.

En raison de sa simplicité, sa rapidité de calcul, et de son pouvoir discriminant, le LBP a été utilisé pour résoudre différents problèmes d'analyse de textures [94] et de visages comme la détection de visages [160] et la reconnaissance de visages [4, 5]. Plusieurs extensions du principe original du descripteur LBP sont ainsi introduites comme les LBP uniformes [104], les LTP (*Local Ternary Patterns*) [130] ou les LBP spatio-temporels (*Volume Local Binary Pattern, VLBP*) [162].

3.5.2 LBP pour la reconnaissance de visages 2D

L'application des LBP à la reconnaissance de visages a été introduite par Ahonen *et al.* [4, 5]. La procédure consiste à utiliser les LBP pour construire plusieurs descriptions locales du visage, et à les combiner en une description globale. La démarche a été motivée par le fait que les méthodes basées sur les caractéristiques locales et/ou hybrides semblent être plus robustes aux variations de poses ou d'éclairage que les méthodes holistiques. L'image de visage est divisée en régions locales (généralement à l'aide d'une grille) et les descripteurs LBP sont extraits de chaque région de manière indépendante. Pour chaque région, un histogramme est calculé en accumulant dans des *bins* chacun des codes LBP extraits. Tous les histogrammes sont ensuite concaténés pour construire un histogramme final qui encode les relations d'apparence et spatiales entre les régions du visage, et qui donne une représentation globale tout en gardant une description locale de chaque région. La Figure 3.14 montre un exemple de calcul des LBP pour représenter un visage 2D. Dans leurs expérimentations,

Ahonen *et al.* ont comparé les LBP à d'autres approches comme EBGM (*Elastic Bunch Graph Matching*) [151] et *EigenFaces* [139]. Les résultats montrent que la méthode basée sur les LBP est invariante aux changements d'expressions et qu'elle est plus robuste aux changements d'éclairage que les autres méthodes.

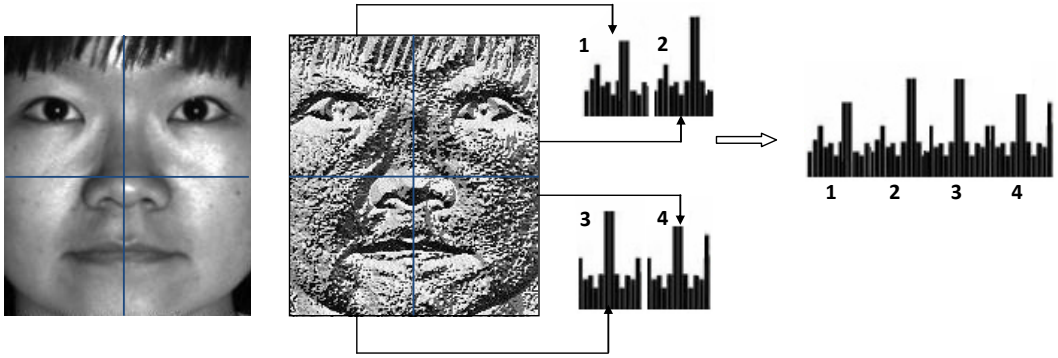

FIGURE 3.14: Exemple de calcul d'un histogramme de LBP sur une image de visage.

Les LBP possèdent plusieurs propriétés qui favorisent leur utilisation pour la description du visage comme :

- la taille du vecteur caractéristique est compacte. Pour un voisinage $V = 8$ et un rayon $R = 1$, seulement 256 codes sont utilisés pour représenter la texture d'une image ;
- son principe est simple tout en offrant un pouvoir discriminant très élevé ;
- les caractéristiques sont très rapides à calculer et ne nécessitent pas beaucoup de paramètres à régler ;
- le code LBP est invariant aux changements linéaires d'éclairage. Ceci est dû à la tolérance aux variations monotones des niveaux de gris. Par conséquent, aucune normalisation de l'éclairage n'est nécessaire – sous l'hypothèse que le changement d'éclairage est linéaire.

3.5.3 Extensions aux visages 3D

L'efficacité du descripteur LBP dans la représentation de visages 2D a amené les chercheurs à l'utiliser sur des images de profondeur. En effet, la définition des LBP permet également de décrire une information de profondeur puisque, comme pour la texture, la profondeur d'un point du visage est fortement liée à son voisinage. Selon sa définition [5], LBP peut également décrire les variations locales de la forme. Lors de l'application des LBP sur une image de profondeur, chacun des 2^V codes LBP représente un motif 3D comme une forme plane, concave ou convexe, au lieu de contours, spots lumineux ou sombres, etc. des LBP classiques (voir la Figure 3.15).

Ainsi, de nombreux travaux récents ont proposé l'application des descripteurs LBP pour la représentation des images de profondeur [64, 91, 133, 145, 155]. Dans [64], une méthode bimodale de reconnaissance de visages a été proposée. Le descripteur LBP a été utilisé afin

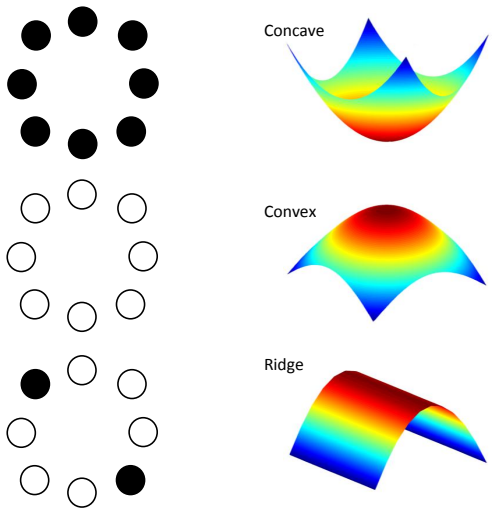

FIGURE 3.15: Exemples de formes 3D pouvant être représentées par $LBP_{(1,8)}$.

d'extraire les caractéristiques 2D et 3D du visage. Les auteurs ont montré que le taux de reconnaissance obtenu en utilisant LBP est meilleur que celui obtenu avec les images de profondeur originales. De la même manière, Xiong *et al.* [155] ont proposé une approche bimodale de reconnaissance de visages, où le visage est représenté en concaténant deux vecteurs caractéristiques (stratégie de fusion précoce) obtenus par extraction des LBP sur l'image d'intensité et l'image de profondeur. Dans Tang *et al.* [133], les auteurs proposent d'appliquer le descripteur LBP sur un maillage 3D afin d'extraire le vecteur caractéristique de forme. Dans leur approche, deux vecteurs caractéristiques sont calculés, en appliquant le principe des LBP sur les valeurs de profondeur et sur les valeurs des vecteurs normaux. Les résultats montrent que l'application des LBP sur les normales donne des précisions plus élevées qu'avec les valeurs de profondeur. La fusion des deux vecteurs (stratégie précoce) donne les meilleurs taux de reconnaissance. Dans [145], Wang *et al.* ont montré que la fusion des LBP avec d'autres descripteurs tels que les filtres de Gabor ou les composantes principales de l'image de profondeur donne de meilleurs résultats que ceux obtenus en utilisant chacun de ces descripteurs séparément. Li *et al.* [91] ont proposé de fusionner les descripteurs LBP avec un autre descripteur d'orientation afin d'améliorer la précision de la reconnaissance de visages basée sur les cartes de profondeur. Une extension multi-échelles (*Multi-Scale LBP, MS-LBP*) a été proposée dans [68], dans laquelle différentes combinaisons de paramètres de rayon et de voisinage ont été utilisées. Dans leurs résultats, les auteurs ont montré que l'utilisation de cette extension multi-échelles avec 4 rayons différents ($R \in \{3,4,5,6\}$) donne de meilleurs résultats par rapport à l'utilisation du descripteur LBP original.

Cependant, concernant l'application directe des LBP sur les images de profondeur, il est à noter qu'elle est susceptible d'entraîner une certaine confusion qui peut diminuer son pouvoir de discrimination. Cette confusion est due au fait d'attribuer le même code LBP pour des formes similaires, mais qui ont des magnitudes différentes (voir la Figure 3.16). En effet, cette différence de magnitude entre les formes est une information importante pour la discrimination, qui est perdue dans l'encodage. Cette perte d'information provient du mode de

fonctionnement des LBP classiques, qui considère seulement le signe de la différence entre le pixel et son voisinage. Bien que les LBP soient indiqués pour l'analyse de visages 2D, leurs application directe n'est pas suffisante pour représenter les images de profondeur, où les différences en profondeur sont nécessaires pour pouvoir distinguer des formes similaires.

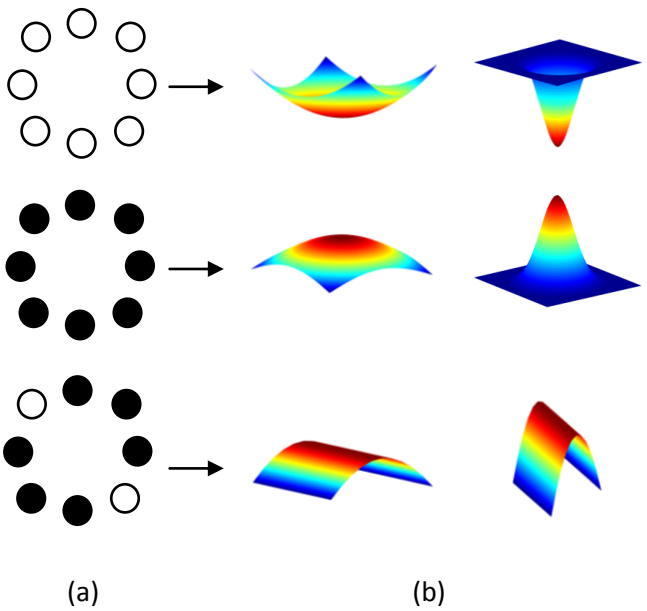

FIGURE 3.16: Confusion entre des formes 3D similaires. (a) Les codes LBP. (b) Les formes 3D possibles correspondantes.

Afin de résoudre ce problème, Yonggang Huang *et al.* [69] ont proposé une version étendue des LBP, nommée 3DLBP. Outre les informations fournies par LBP, 3DLBP considère également la valeur de la différence de profondeur notée *DD* (*Depth Difference*) entre le pixel central et son pixel voisin. Un ensemble de trois bits est utilisé pour coder les valeurs absolues des différences de profondeur. Il est ensuite combiné aux signes (0 ou 1) obtenus par les LBP. En se basant sur une étude statistique, les auteurs considèrent une valeur maximale de 7 pour ces valeurs de différence. En effet, dans leur étude, les auteurs ont noté que plus de 93% des *DD* obtenues entre les pixels avec un rayon $R = 2$ sont inférieures à 7. Ceci est dû à l'aspect lisse de la profondeur du visage, ce qui n'est pas le cas pour les images en niveaux de gris où les valeurs de différence entre les pixels voisins peuvent être arbitrairement grandes selon la texture et les conditions de l'environnement. Une *DD* est donc représentée par quatre bits i_1, i_2, i_3, i_4. Le bit i_1 est le code représentant le signe, et i_2, i_3, i_4 sont les trois bits représentants la valeur absolue de la différence de profondeur ($|DD|$). Toute différence supérieure au seuil est ramenée à 7. Les 4 bits sont ensuite séparés en 4 couches. Pour chaque couche, les bits correspondants de l'ensemble des *DD* des pixels voisins sont concaténés pour générer un code binaire. Au total, pour chaque pixel, on obtient 4 codes binaires P_1, P_2, P_3, P_4, qui sont ensuite convertis en 4 valeurs décimales. L'histogramme de chaque couche est calculé, puis les histogrammes sont concaténés pour former un descripteur unique pour l'image. La Figure 3.17 illustre le mode de fonctionnement du 3DLBP.

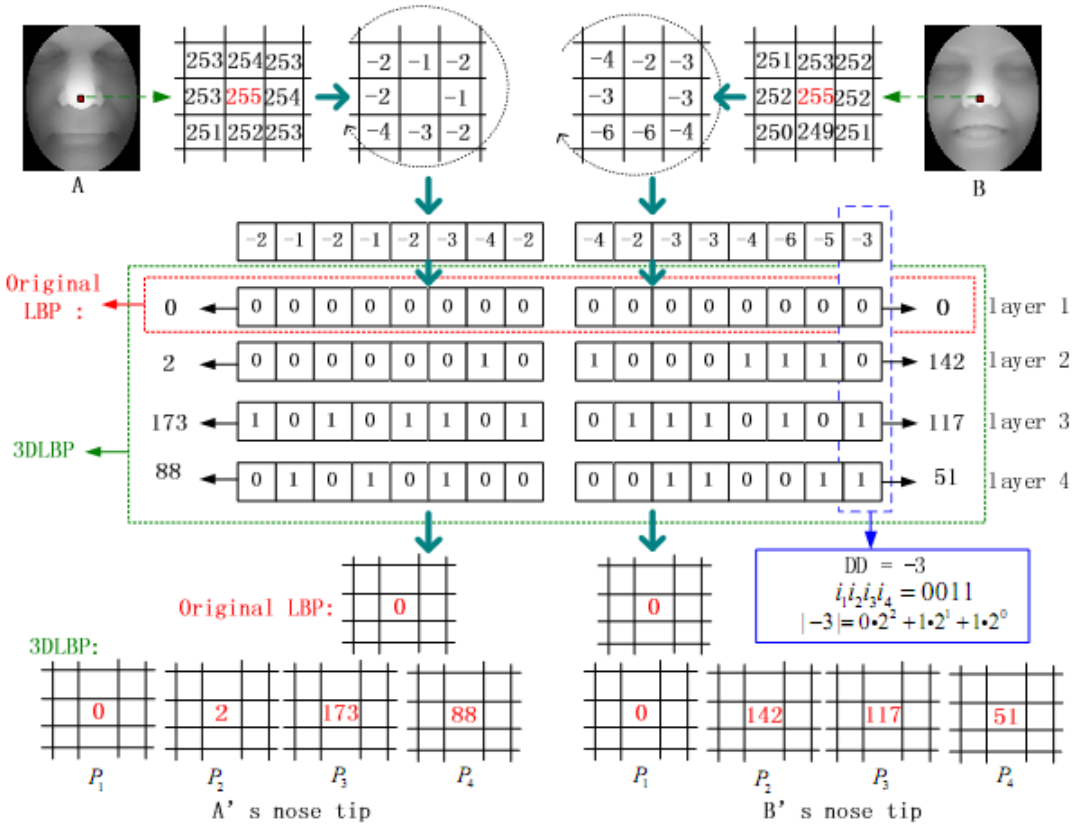

FIGURE 3.17: Exemple de calcul des 3DLBP, extrait de [69].

Les auteurs ont évalué le descripteur proposé sur la collection FRGC [110]. Les résultats montrent que la précision obtenue en utilisant les 3DLBP est meilleure que celle des LBP et des cartes de profondeur brutes. Bien que l'utilisation des 3DLBP augmente le pouvoir discriminant des LBP en utilisant, en plus du signe, la valeur de différence de profondeur dans le voisinage, ce descripteur souffre de certaines limites :

- la taille du vecteur caractéristique est très grande. Chaque image est représentée par 4 matrices de même taille. Ces matrices sont segmentées en régions et un histogramme par région pour chaque matrice est calculé. Ce qui donne un vecteur caractéristique de taille $H * 4 * 256$, où H est le nombre de régions de découpage. Ceci implique un temps de traitement plus long ;
- la méthode de codage n'est pas consistante avec la variation de la profondeur. Une très faible variation dans le voisinage du pixel peut conduire à une grande différence dans les codes calculés. Par exemple, pour un pixel de valeur 10 et dont le voisinage est $\{13,14,15,12,11,9,11,12\}$, les différences de profondeur sont alors $\{3,4,5,2,1,-1,1,2\}$, les codes 3DLBP obtenus sont $\{251,96,145,174\}$. Pour un changement mineur de la valeur du premier voisin de 13 vers 14, on obtient un code 3DLBP $\{251,224,17,46\}$. Ceci est dû à la manière de répartir les valeurs des 4 bits sur les différentes couches encodant les DD ;

– les valeurs *DD* sont considérées inférieures ou égales à 7 et sont codées par conséquent sur 3 bits. Cependant, l'étude statistique sur laquelle les auteurs se basent est limitée à un petit rayon ($R = 2$). Par conséquent une extension multi-échelles ne peut pas être envisagée en utilisant les 3DLBP, puisque le seuil de 7 pour les *DD* ne serait alors plus valide. Afin de généraliser cette approche pour qu'elle puisse être utilisée avec différents rayons, il est nécessaire de représenter la différence sur 8 bits, soit 256 différences de profondeur possibles. Cependant, cette généralisation a le défaut d'engendrer une représentation de taille importante : 9 matrices pour représenter un seul visage de profondeur.

Comme la stratégie multi-échelles a montré de meilleurs résultats que les LBP originaux dans la reconnaissance 2D, Di Huang [1] *et al.* [65] ont proposé une version multi-échelles des 3DLBP proposée par Yonggang Huang *et al.* [69], où les réponses de 3DLBP à différentes échelles sont combinées afin de fournir une représentation plus efficace des visages de profondeur. À la différence du travail de Yonggang Huang *et al.* [69] dans lequel des histogrammes sont calculés à partir des matrices obtenues par le 3DLBP, les auteurs dans [65] ont utilisé directement les matrices de codes, appelées *MS-eLBP-DFs (Multi-Scale extended Local Binary Pattern Depth Faces)*, afin de conserver l'information spatiale (voir la Figure 3.18). Les résultats obtenus dans leurs travaux pour différentes valeurs de rayons (2 à 8) sont presque identiques. Ceci est dû au mode de fonctionnement du 3DLBP : comme ce dernier est basé sur une étude statistique limitée à un petit rayon ($R = 2$), une grande différence de profondeur résultante d'un grand rayon $(R > 2)$ est systématiquement ramenée à 7, et n'est donc pas prise en compte. L'information de profondeur à grande échelle n'a donc pas du tout été exploitée dans le travail proposé par les auteurs, ce qui explique la similarité des résultats obtenus à différentes échelles.

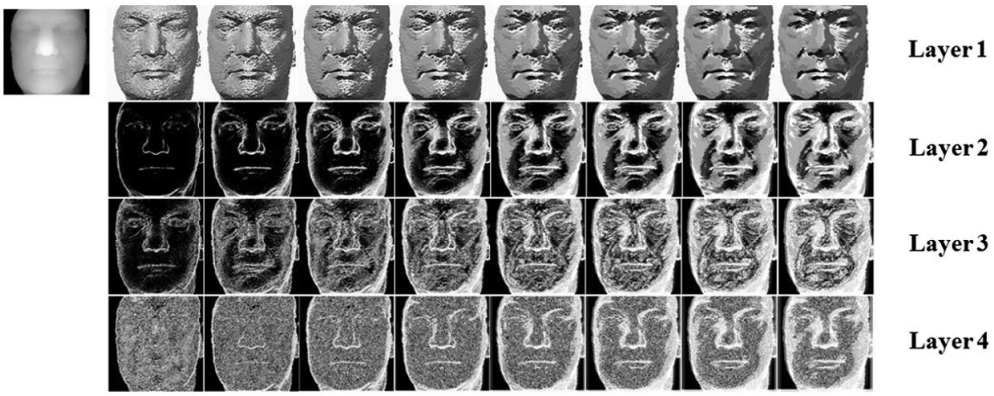

FIGURE 3.18: Exemple des *MS-eLBP-DFs* utilisées pour la reconnaissance 3D par Di Huang *et al.* [65].

1. Les auteurs Yonggang Huang [69] et Di Huang [65] ont le même nom. Pour cette raison, nous citons également leurs prénoms.

3.6 Conclusion et positionnement

Dans ce chapitre, nous avons donné un aperçu global des approches de reconnaissance faciale. Une taxonomie complète a été présentée en classant les approches de l'état de l'art en fonction de la modalité de reconnaissance utilisée : 2D, 3D ou bimodale 2D-3D, et en analysant les points forts et faibles pour chaque modalité. Les points faibles de la reconnaissance 2D, notamment la sensibilité aux variations d'éclairages et de la pose, peuvent être compensés par l'information de forme 3D invariante à ces changements. Les méthodes bimodales 2D-3D ont démontré un gain notable de précision par rapport à l'utilisation de chaque modalité séparément. Beaucoup de méthodes récentes visent donc à utiliser les deux modalités pour la reconnaissance de visages. La majorité de ces méthodes se basent sur une fusion tardive (fusion de décisions) en considérant que les deux modalités sont indépendantes, du fait que chacune fournit un aspect différent du visage. Cependant, il existe une certaine complémentarité entre les données visuelles et les données de profondeur, qui peut être exploitée en suivant une stratégie de fusion précoce (fusion de données brutes ou de descripteurs). En effet, les positions des traits de visages (nez, yeux, etc.) ou les éventuelles occultations sont les mêmes dans les deux types de données. L'hypothèse que les mesures 2D et 3D sont indépendantes peut donc engendrer une perte d'information, et ainsi entraver la synergie des deux modalités. Au meilleur de notre connaissance, aucun consensus ne se dégage des approches étudiées quant à une meilleure stratégie de fusion valable universellement. Nous proposons ainsi une nouvelle stratégie bi-niveaux qui permet de combiner ces stratégies afin de bénéficier des avantages de la fusion précoce et de la fusion tardive.

La description des images de profondeur a également été étudiée dans ce chapitre. Nous nous somme intéressés à l'application des LBP sur les images de profondeur. En effet, le succès de ce descripteur à motivé les chercheurs à étendre son principe pour décrire les données de profondeur. L'extension principale proposée dans l'état de l'art est le 3DLBP qui consiste en l'utilisation de la magnitude en plus du signe des différences de profondeur afin de représenter les formes 3D plus finement. Le descripteur 3DLBP a donné des résultats très prometteurs. Cependant, des aspects comme la grande taille du vecteur descripteur, les limitations multi-échelles et la très grande sensibilité aux petits changements de profondeur constituent les grands problèmes de cette méthode. Nous proposons une extension alternative aux LBP, sous la forme d'un nouveau descripteur qui intègre l'information de la magnitude, et dont le mode de codage est flexible et applicable à différentes échelles. Nos propositions sont entièrement présentées dans la partie suivante.

Deuxième partie

Approche proposée pour la reconstruction et la reconnaissance de visages

Chapitre 4

Aperçu global de l'approche bimodale 2D-3D de reconnaissance de visages

Sommaire

4.1 Introduction 58
4.2 Acquisition 58
4.3 Reconnaissance 59
4.4 Fusion 60
4.5 Conclusion 60

4.1 Introduction

Après avoir donné dans la partie précédente un aperçu global de l'état de l'art des deux volets principaux de cette thèse (acquisition 3D et reconnaissance de visages), nous introduisons dans cette partie du manuscrit l'approche globale proposée pour la reconnaissance bimodale 2D-3D de visages. Dans ce chapitre nous donnons une description globale de l'approche proposée ainsi que des modules principaux aux niveaux desquels nous avons apporté des contributions. La Figure 4.1 illustre les différents composants du système de reconnaissance proposé (les contributions sont désignées par une étoile). Dans la suite de ce chapitre, nous décrivons chacun de ces modules.

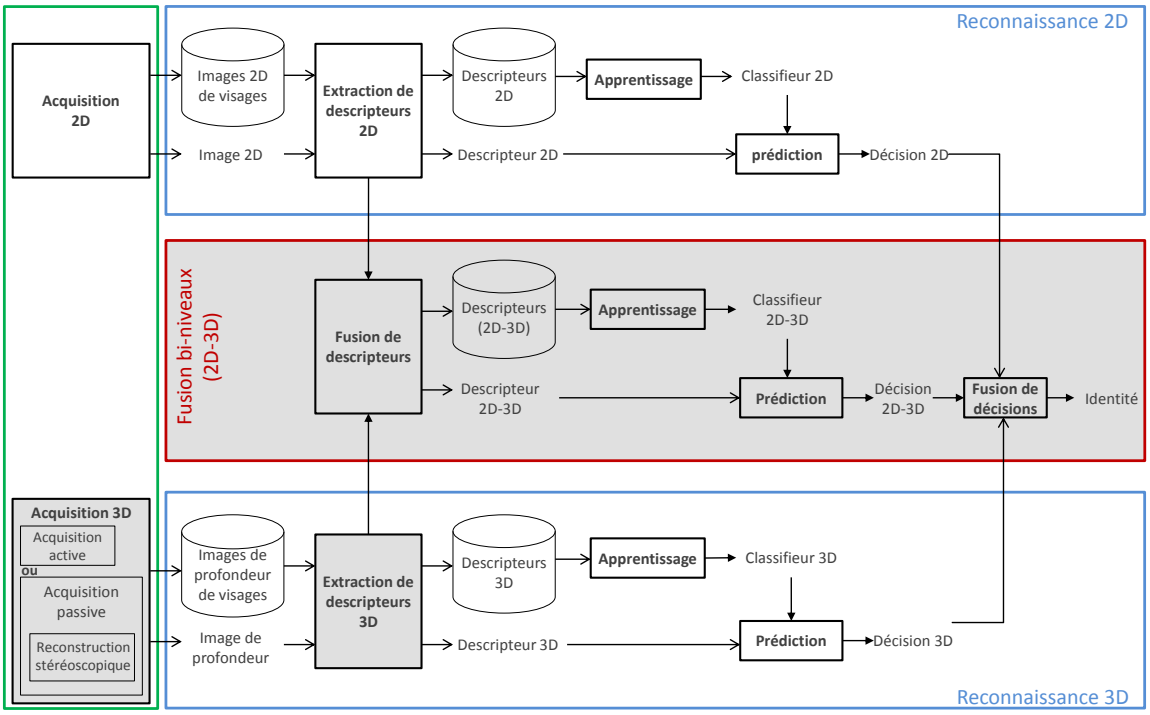

FIGURE 4.1: Schéma global de l'approche proposée. Les boîtes grisées représentent nos contributions.

4.2 Acquisition

Ce module permet d'acquérir des informations sur le visage de la personne à reconnaître. Dans notre approche, le visage est représenté par deux types de données :

- les données 2D : ce sont des images 2D acquises simplement avec des appareils photographiques. Nous notons que les images utilisées sont en niveaux de gris ;
- les données 3D : elles sont représentées par des images de profondeur. Comme le montre le schéma de la Figure 4.1, les données 3D peuvent être acquises directe-

ment par des méthodes actives (scanner 3D, caméra de type Kinect, etc.) ou par des méthodes de reconstruction passives (caméra stéréoscopique). Dans le cadre de cette thèse, nous nous sommes intéressés aux différents types d'acquisition. Une méthode pour la **reconstruction stéréoscopique du visage** est proposée. Elle permet d'estimer la carte de profondeur du visage à partir de deux images stéréoscopiques en se basant sur la méthode d'appariement de *block matching* et les propriétés topologiques du visage. Cette méthode est présentée dans le chapitre 5. L'acquisition de la profondeur à partir de Kinect a été aussi effectuée dans le cadre de construction d'une collection de tests bimodale 2D-3D. Aucune étape de reconstruction n'est nécessaire car le capteur fournit directement l'image 2D et la carte de profondeur correspondante. Cette partie est décrite dans le chapitre 7. Les données 3D obtenues à partir de scanners 3D mises à disposition sous la forme de collections de tests publiques sont aussi utilisées.

4.3 Reconnaissance

Une fois les données des deux modalités acquises, le processus de reconnaissance est effectué. Celui-ci consiste d'abord en **l'extraction de descripteurs** suivie par la **classification (apprentissage et prédiction)**.

- **Extraction de descripteurs** : cette étape consiste à extraire les vecteurs caractéristiques à partir des images 2D et des images de profondeur obtenues préalablement. Ces vecteurs contiennent des valeurs discriminantes qui caractérisent le visage d'une personne donnée par rapport aux autres. Ceci est effectué pour les données des deux modalités en utilisant deux descripteurs :
  - pour la représentation des images 2D, nous utilisons une méthode de l'état de l'art basée sur le descripteur LBP. Ce descripteur a connu beaucoup de succès dans l'état de l'art grâce à son pouvoir discriminant et sa simplicité, ce qui a motivé notre choix ;
  - par ailleurs, une de nos contributions réside dans la description des données 3D du visage. Nous proposons un descripteur 3D inspiré du principe du LBP classique. Cette extension prend en considération la nature des images de profondeur, permettant ainsi une meilleure représentation de celles-ci par rapport à l'application directe des LBP conçus à l'origine pour les images 2D. Ce descripteur est présenté dans le chapitre 6.
- **Classification (apprentissage et prédiction)** : les vecteurs caractéristiques, obtenus après l'étape d'extractions de descripteur, sont utilisés dans le processus de classification. Ce dernier consiste, dans un premier temps, à appliquer un processus d'apprentissage sur les vecteurs caractéristiques (étiquetés selon l'identité du visage) d'une collection de visages afin de construire un modèle de reconnaissance (un classifieur). Dans un deuxième temps, une étape de prédiction est effectuée afin d'affecter une identité à un visage inconnu donné en utilisant le classifieur construit par l'étape d'apprentissage. Le processus de classification (apprentissage et prédiction) est effectué de la même manière pour les deux modalités 2D et 3D.

4.4 Fusion

À la lumière de ce qui a été présenté dans l'état de l'art des approches bimodales 2D-3D, nous proposons de combiner les données 2D et 3D du visage en utilisant une stratégie qui permet d'exploiter la relation entre les deux modalités avant et après la classification. Nous proposons donc une **stratégie bi-niveaux** afin d'effectuer la reconnaissance bimodale. Celle-ci permet de tirer profit des deux stratégies de fusion (la fusion précoce et la fusion tardive).

Notre approche utilise trois ensembles de vecteurs caractéristiques pour représenter les visages. Les deux premiers sont obtenus à partir des deux modalités 2D et 3D comme nous avons expliqué ci-dessus. Le troisième est obtenu en effectuant une concaténation de ces deux vecteurs (fusion précoce). Il contient donc les données des deux modalités. Lors de l'apprentissage, trois classifieurs sont ainsi construits indépendamment à partir des trois ensembles de descripteurs. Lors de la prédiction de l'identité d'un visage donné, trois vecteurs caractéristiques sont donc nécessaires : un vecteur 2D, un vecteur 3D et un vecteur 2D-3D. Trois processus de prédiction sont mis en œuvre en utilisant ces vecteurs et les trois classifieurs correspondants, construits lors de l'étape d'apprentissage. Une fusion tardive est ensuite proposée pour combiner les trois décisions obtenues.

4.5 Conclusion

Le but de ce court chapitre est de donner une vue globale des différentes contributions proposées dans le cadre de ce travail, ainsi que les relations entre elles. La reconnaissance bimodale 2D-3D peut être effectuée en utilisant différents types de capteurs. En effet, l'acquisition 3D via la stéréovision est une problématique indépendante de la reconnaissance. Néanmoins, la reconstruction stéréoscopique est une alternative prometteuse d'acquisition de données pour les approches bimodales de reconnaissance de visages. Dans le deuxième chapitre de cette partie, nous nous concentrons sur la première contribution de ce travail de thèse, l'acquisition 3D, où nous présentons la méthode de reconstruction stéréoscopique de visages proposée. Ensuite, la question de la reconnaissance est traitée dans le troisième chapitre, où nous décrivons le descripteur proposé pour la représentation des images de profondeur, ainsi que la stratégie de fusion permettant de combiner les deux modalités en vue de proposer une approche bimodale de reconnaissance de visages.

Chapitre 5

Reconstruction 3D du visage

Sommaire

5.1 Introduction 62
5.2 Reconstruction stéréoscopique basée sur la structure topologique du visage 62
5.2.1 Construction du modèle de disparité 62
5.2.2 Calcul de la carte de disparité 63
5.3 Post-traitement : débruitage de la carte de profondeur 65
5.3.1 Détection du bruit 66
5.3.2 Suppression du bruit 68
5.4 Conclusion 68

5.1 Introduction

Dans ce chapitre, nous présentons la méthode proposée pour la reconstruction faciale stéréoscopique. L'originalité de notre méthode réside dans l'intégration d'a priori sur la forme du visage dans le processus de reconstruction pour améliorer le résultat d'appariement stéréoscopique. Nous proposons aussi un algorithme de post-traitement afin d'éliminer les artefacts éventuellement présents dans la carte de profondeur reconstruite.

5.2 Reconstruction stéréoscopique basée sur la structure topologique du visage

Afin d'estimer la carte de disparité, les mesures basées sur l'intensité utilisées dans les méthodes d'appariement stéréoscopique ne sont pas toujours suffisantes pour récupérer précisément la géométrie 3D, en particulier dans les régions peu texturées de la scène (cf. problème d'ouverture décrit dans la Section 2.4.2). Comme le visage est un objet spécifique ayant une structure et des propriétés propres, il peut donc être utile d'intégrer des a priori sur ses propriétés afin de guider le processus de reconstruction vers les caractéristiques désirées, tout en gardant une complexité de calcul faible.

Pour reconstruire la forme 3D du visage, nous proposons de construire un modèle épars de disparité du visage à partir des images stéréoscopiques en se basant sur quelques a priori comme l'aspect lisse de sa surface, ainsi que sa structure topologique ; un visage est formé par ensemble de composants (deux yeux, un nez, une bouche, etc.) positionnés selon un certain ordre au long de l'axe de profondeur. Ce modèle donne une représentation globale de la disparité des points du visage qui serviront de guide lors de l'étape de calcul de la carte de disparité. Le modèle de disparité est construit indépendamment du processus d'appariement stéréoscopique et utilisé pour l'initialisation de ce dernier. Le processus d'appariement est effectué en tenant compte des informations topologiques obtenues par le modèle de disparité et de la propriété de l'aspect lisse du visage afin d'estimer la carte de disparité dense. Pour l'appariement stéréoscopique, nous avons choisi d'utiliser la méthode de *block matching* (BM) pour sa simplicité et sa rapidité.

5.2.1 Construction du modèle de disparité

La première étape de notre méthode consiste à construire un modèle épars de disparité du visage pour lequel nous cherchons à estimer la carte de disparité. Afin de construire ce modèle, nous commençons par un processus d'ajustement d'un modèle actif de forme (*Active Shape Model, ASM*) [101] sur les deux images (gauche et droite) indépendamment pour localiser un ensemble de points mis en correspondance avec une confiance élevée.

L'ASM est un modèle statistique de forme obtenu à partir d'un processus d'apprentissage supervisé sur une collection de visages annotés (cf. Section 3.2.3). Nous utilisons le processus d'ajustement de l'ASM parce qu'il utilise les informations de la structure topologique du visage, ce qui garantit une bonne localisation des points caractéristiques du visage dans

la paire stéréoscopique, et par conséquent une confiance élevée pour l'appariement de ces points.

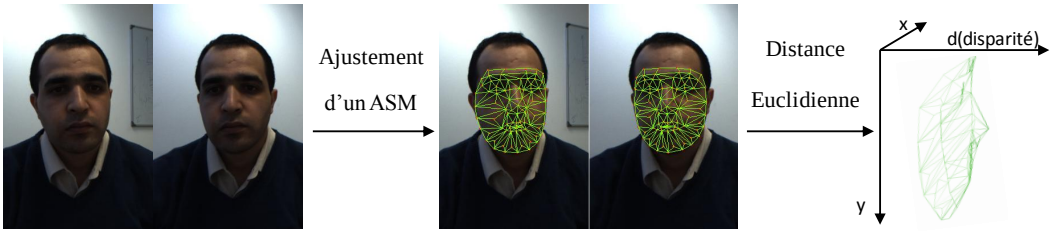

FIGURE 5.1: Construction du modèle de disparité.

Après l'ajustement de l'ASM sur les images droite et gauche séparément, nous obtenons les coordonnées 2D des n points caractéristiques dans l'image droite $R = \{(x_i, y_i) | i \in [1,n]\}$ et dans l'image gauche $L = \{(x_i^{'}, y_i^{'}) | i \in [1,n]\}$, qui sont ensuite utilisées pour obtenir l'ensemble de coordonnées $3D$ $P = \{p_i(x,y,d) | i \in [1,n]\}$ qui représente le modèle de disparité du visage concerné (voir la Figure 5.1). La disparité d des points est calculée par la distance euclidienne de la manière suivante :

$$d_i = \sqrt{(x_i - x_i^{'})^2 + (y_i - y_i^{'})^2} \tag{5.1}$$

Une fois le système calibré et les paires stéréoscopiques rectifiées, les coordonnées y des points correspondants sont identiques : ces points sont sur une même ligne horizontale. Par conséquent, la disparité est calculée en utilisant seulement les coordonnées x de la manière suivante :

$$d_i = \sqrt{(x_i - x_i^{'})^2} = |x_i - x_i^{'}| \tag{5.2}$$

Dans le cas où les coordonnées y des points correspondants résultant du processus d'ajustement ne sont pas identiques suite à une erreur d'ajustement (causée généralement par la présence du bruit dans les images) nous appliquons une normalisation à la moyenne. Le modèle de disparité du visage construit à cette étape va servir de guide dans l'étape de calcul de la carte de disparité dense que nous allons présenter dans la suite.

5.2.2 Calcul de la carte de disparité

La carte de disparité dense est calculée en deux étapes. Dans la première étape, un processus consistant à décomposer le modèle de disparité du visage en un ensemble de *plans de niveaux* est effectué (voir la Figure 5.2). Les plans de niveaux sont définis perpendiculairement à l'axe de la profondeur. Ils sont donc parallèles au plan image. Les valeurs de disparité du modèle épars sont utilisées afin de construire ces plans. Le plan le plus éloigné correspond à la disparité minimale du modèle. Le plan le plus proche correspond à la disparité maximale du modèle (ce point est généralement le bout du nez lorsque le visage est sous une pose frontale ou si l'angle de prise de vue est petit). Nous définissant trois autres plans

fixés en prenant les trois disparités ayant le maximum d'occurrences. Ainsi nous utilisons 5 plans au total.

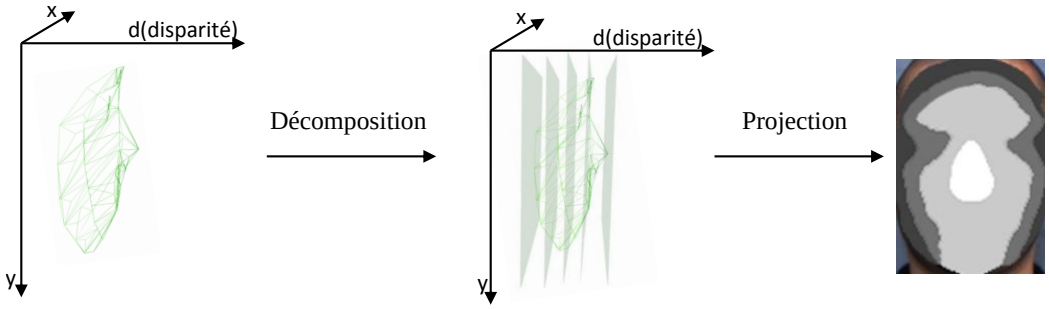

FIGURE 5.2: Décomposition du modèle de disparité.

Après l'étape de décomposition, nous pouvons définir différentes régions dans l'image de visage qui correspondent à différents intervalles de disparité (situés entre deux plans de niveau). L'intervalle de disparité attribué à chaque région est donc défini selon le plan de niveau auquel les points limitant cette région appartiennent (voir la Figure 5.2). Pour une région r_i, nous définissons donc l'intervalle de disparité comme : $[\text{DispMin}_i, \text{DispMax}_i[$ où DispMin_i et DispMax_i sont les valeurs de disparité associées aux deux plans limitant la région r_i. Dans la Figure 5.3, nous montrons différentes étapes de décomposition du modèle de disparité selon la pose du visage.

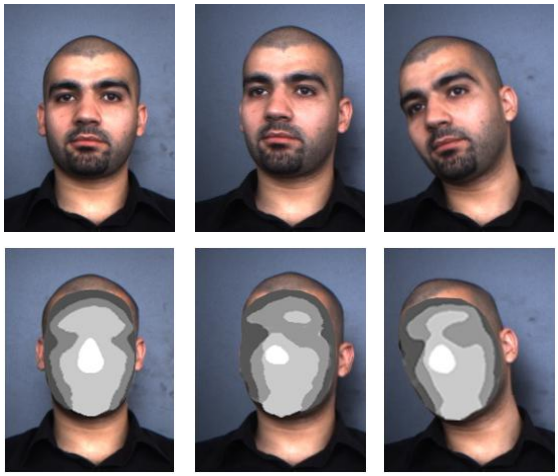

FIGURE 5.3: Différentes projections du modèle de disparité pour un visage sous différentes poses.

L'étape de décomposition associe un intervalle de disparité à chaque région du visage. Ceci réduit la zone de recherche lors de l'appariement : la recherche d'un pixel correspondant dans toute la ligne épipolaire est réduite à seulement un petit segment défini par l'intervalle associé à chaque région. Elle permet également de garantir une certaine régularité de la carte

finale de disparité estimée, ainsi qu'une réduction du nombre de pics. Ceci est dû au fait qu'elle limite l'intervalle de disparité de chaque zone du visage au minimum et au maximum des intervalles des zones du voisinage.

Dans la deuxième étape, nous calculons la disparité de tous les points du visage en utilisant l'algorithme de *block matching*. Étant donnés un pixel $p_r(x,y)$ appartenant à la région r_i dans l'image droite, une fenêtre d'appariement w de taille $n \times m$, et un intervalle de disparité $[\text{DispMin}_i, \text{DispMax}_i[$ attribué à la région r_i, l'objectif est d'obtenir la disparité $d \in [\text{DispMin}_i, \text{DispMax}_i[$, qui donne la meilleure correspondance entre le pixel $p_r(x,y)$ et un pixel de l'image gauche $p_l(x+d,y)$ selon le critère de mise en correspondance choisi. Dans le cadre de notre travail, nous avons utilisé la somme des différences absolues (*Sum of Absolute Differences, SAD*) [62] définie par :

$$SAD_{(I_l(x,y),I_r(x',y'))} = \sum_{u=0}^{m-1}\sum_{v=0}^{n-1} |I_l(x+u,y+v) - I_r(x'+u+d,y'+v)| \tag{5.3}$$

où :

- I_l (resp. I_r) est l'image gauche (resp. droite) ;
- $m \times n$ est la taille de la fenêtre d'appariement w.

Enfin, en utilisant la carte de disparité, la carte de profondeur est calculée selon l'équation suivante :

$$z = \frac{fb}{d} \tag{5.4}$$

où f est la focale de la caméra et b est la *baseline* du système stéréoscopique (cf. Section 2.4.1).

5.3 Post-traitement : débruitage de la carte de profondeur

La carte de profondeur estimée contient souvent des artefacts (trous ou bruit) en raison des données manquantes ou erronées obtenues lors du processus de mise en correspondance (cf. Section 2.4.4). Bien que la méthode d'estimation de disparité proposée vise à minimiser le bruit, une étape de débruitage reste indispensable pour supprimer le bruit éventuellement présent dans la carte de profondeur. Contrairement aux méthodes de débruitage qui affectent la totalité de l'image, nous avons choisi de suivre une méthodologie locale afin d'éviter toute perte de données. Elle consiste en deux étapes : la détection du bruit et la suppression de bruit. En d'appuyant sur l'hypothèse du caractère lisse de la profondeur du visage, nous proposons une méthode traitant le problème de détection des zones bruitées dans une carte de profondeur. Cette hypothèse a été montrée par les travaux statistiques de Huang *et al.* [69]. Les auteurs affirment que les différences de profondeur entre les pixels voisins sont inférieures à 7 pour 93% des pixels de la carte de profondeur des visages.

5.3.1 Détection du bruit

La détection de bruit est une étape essentielle qui permet de définir les données erronées dans la carte de profondeur. Elle est facilement réalisable pour la détection de bruit de petite taille (en forme de pics). Cependant, elle devient plus complexe quand il s'agit de détecter les zones bruitées de grande taille. Nous proposons une méthode pour identifier automatiquement le bruit dans la carte de profondeur du visage en intégrant la propriété de l'aspect lisse de sa forme 3D. Une méthode consistant en une analyse locale et globale est utilisée pour segmenter la carte de profondeur selon deux classes : zones bruitées et zone non-bruitées. Après l'identification, la suppression du bruit (modification des données erronées) peut être effectuée en utilisant n'importe quelle méthode de l'état de l'art.

Le processus consiste à rechercher les zones bruitées en parcourant la carte de profondeur. Comme la surface du visage est lisse horizontalement et verticalement, nous pouvons traiter les données indifféremment au niveau des lignes ou des colonnes. Dans ce travail, nous choisissons arbitrairement les lignes.

La détection du bruit est effectuée en deux étapes : segmentation et classification. Tout d'abord, chaque ligne de profondeur est découpée en différents segments en se basant sur la dérivée de la profondeur. Ensuite, les segments sont classés en segments bruités ou non-bruités en analysant localement et globalement la ligne de profondeur.

L'étape de segmentation consiste à diviser chaque ligne de profondeur en un ensemble de segments. Nous considérons la profondeur comme étant une fonction f définie de la manière suivante :

$$\begin{aligned} f: \ \mathbb{N}^2 &\rightarrow \mathbb{N} \\ (x,y) &\rightarrow z \end{aligned}$$

où $x \in [0, N-1]$ et $y \in [0, M-1]$ sont les coordonnées des pixels d'une image de profondeur de taille $N \times M$ et z est la valeur de la profondeur à ces coordonnées.

Afin de détecter les principaux points de découpage qui définissent l'ensemble des segments, nous utilisons la dérivée première de la fonction de profondeur f. En effet, considérant l'aspect lisse du visage, la valeur de la dérivée de la fonction f en tout point doit être inférieure à un certain seuil.

Dans ce travail, du fait que nous traitons les lignes de la carte de profondeur séparément, nous limitons la fonction aux coordonnées x uniquement, et nous considérons y comme une constante. Par conséquent, la dérivée de f par rapport à x peut être obtenue numériquement par approximation de la manière suivante :

$$\frac{\partial f(x,y)}{\partial x} \approx \frac{f(x+\delta x, y) - f(x,y)}{(x+\delta x) - x} \tag{5.5}$$

Comme dans notre traitement nous utilisons une image (pixels discrets), nous choisissons $\delta x = 1$ et donc l'équation devient :

$$\frac{\partial f(x,y)}{\partial x} \approx f(x+1, y) - f(x,y) \tag{5.6}$$

Afin de trouver les points de découpage, nous résolvons l'équation ci-dessous :

$$\left|\frac{\partial f(x,y)}{\partial x}\right| \geq t_y \tag{5.7}$$

où t_y est un seuil égal à la moyenne des valeurs de la dérivée calculée en tout x de la ligne y. Il est calculé comme suit :

$$t_y = \frac{\sum_{x=0}^{N-1}\left|\frac{\partial f(x,y)}{\partial x}\right|}{N} \tag{5.8}$$

où N est le nombre de pixels dans la ligne de profondeur.

Dans la Figure 5.4, nous illustrons comment trouver les points de découpage et comment définir les segments principaux pour une ligne de profondeur donnée.

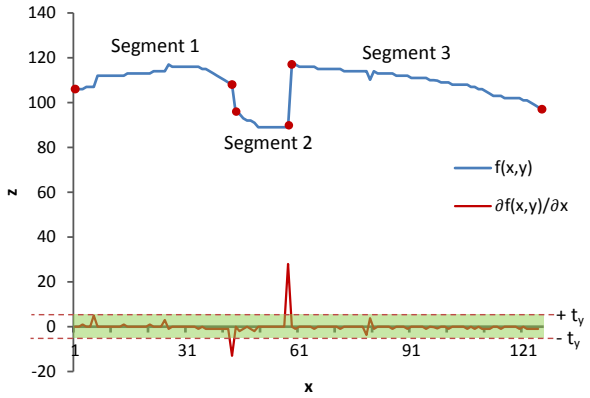

FIGURE 5.4: Détection des points de découpage et décomposition de la ligne de profondeur. Les points rouges désignent les points de découpage.

La seconde étape consiste à identifier les zones de bruit. Dans cette étape, nous allons classer les segments obtenus précédemment pour chaque ligne comme bruités et non-bruités. Pour ce faire, nous utilisons l'écart-type des valeurs de profondeur de toute la ligne, la moyenne de ces valeurs et la moyenne des valeurs de profondeur de chaque segment. Ceci permet de mesurer la dispersion de ce dernier par rapport à la moyenne de toute la ligne de profondeur et donc de l'identifier comme btuité ou non. Dans la Figure 5.5, le segment en rouge est identifié comme étant bruité.

Étant donnée une ligne de profondeur avec une moyenne μ, un écart-type σ et un ensemble de segments $s_1, s_2, ..., s_n$ (obtenus dans l'étape de décomposition) et leurs moyennes correspondantes $m_1, m_2, ..., m_n$, nous identifions un segment s_i de moyenne m_i comme étant bruité si la condition suivante est réalisée : $m_i \notin [\mu - \sigma, \mu + \sigma]$.

Dans la Figure 5.5, nous illustrons le résultat obtenu par l'application du processus de détection de bruit sur une ligne de profondeur bruitée.

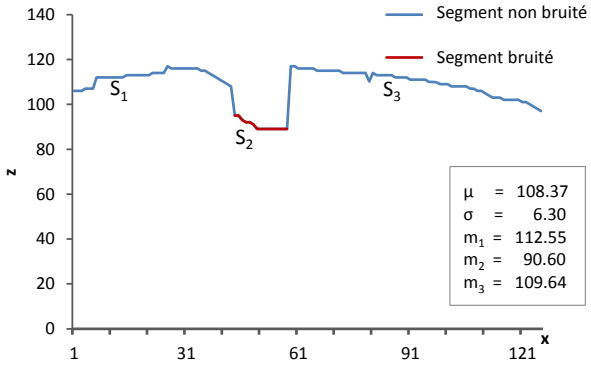

FIGURE 5.5: Détection des segments bruités. Chaque m_i indique la moyenne du segment s_i.

5.3.2 Suppression du bruit

Après la détection du bruit, toutes les zones bruitées sont supprimées et ensuite une étape de remplissage est réalisée en utilisant un algorithme d'interpolation. Pour cela, l'interpolation cubique est utilisée. Dans la Figure 5.6, nous illustrons le processus de débruitage proposé sur un exemple de carte de profondeur.

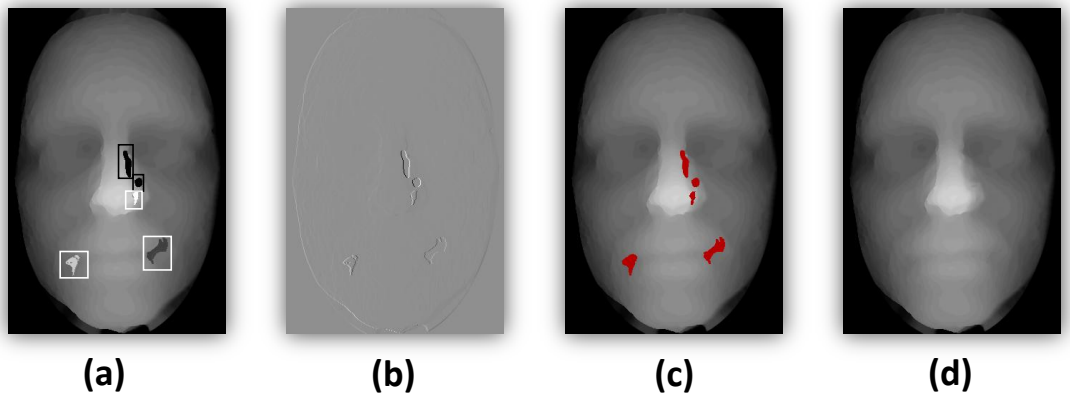

FIGURE 5.6: Débruitage de la carte de profondeur : (a) image bruitée, (b) image des dérivées, (c) détection des segments bruités, (d) image débruitée.

5.4 Conclusion

Après avoir donné un aperçu global résumant les différents modules de l'approche proposée pour la reconnaissance 2D-3D dans le chapitre précédent, nous avons commencé dans ce chapitre par le premier module du processus de reconnaissance : le module d'acquisition. La méthode proposée pour la reconstruction stéréoscopique du visage, qui est une étape principale de l'acquisition 3D par des capteurs stéréoscopiques a été présentée. A la différence

des autres méthodes générales utilisées pour le calcul de la disparité des objets, nous avons introduit une méthode d'estimation de la profondeur spécifique au visage. Elle utilise les caractéristiques de la forme du visage humain (obtenues en ajustant un ASM) afin d'améliorer les résultats des méthodes générales. Notre méthode permet un calcul dense et léger des cartes de profondeur de visages pris de face ou sous petites variations de pose. Cependant, elle ne prend pas en considération les grandes variations de pose. En effet, l'ajustement de l'ASM n'est plus réalisable dans ces conditions. Pour le post-traitement des cartes de profondeur estimées, nous avons adopté une méthode de débruitage local basée sur la détection préalable du bruit. Ceci permet de séparer les données bruitées des données correctement estimées et de les traiter sans affecter le reste des données. Nous avons proposé un algorithme de détection de bruit basé sur la propriété de l'aspect lisse de la surface du visage. Notre algorithme diffère de ceux de l'état de l'art dans sa capacité à identifier les zones bruitées de différentes tailles. Un autre avantage de notre algorithme est qu'il ne nécessite aucun paramétrage et qu'il est entièrement automatique. Notre algorithme est robuste à la taille de la région bruité si elle occupe moins de 50% de la ligne de profondeur. En effet, dans le cas contraire, la classification des segments serait inversée.

Dans le chapitre suivant, nous nous intéressons à l'extraction de descripteurs à partir des cartes de profondeur, ainsi qu'à la fusion bimodale pour la reconnaissance de visages.

Chapitre 6

Reconnaissance bimodale 2D-3D

Sommaire

6.1 Introduction 72
6.2 Extraction des descripteurs 2D et 3D 72
6.3 Descripteur d'images de profondeur 72
6.3.1 Définition du DLBP 73
6.3.2 Stratégie multi-échelles 74
6.3.3 Calcul du seuil 75
6.3.4 Construction des histogrammes 75
6.4 Prédiction de l'identité par fusion 76
6.4.1 Fusion de descripteurs 76
6.4.2 Fusion de décisions 76
6.4.3 Fusion bi-niveaux 78
6.5 Conclusion 78

6.1 Introduction

Après l'étape d'acquisition des données 3D du visage, nous présentons dans ce chapitre les deux parties sur lesquelles s'articulent nos deuxième et troisième contributions : l'extraction de descripteurs et la fusion bimodale pour la reconnaissance de visages. Dans notre approche, un visage est représenté par deux images : une image d'apparence 2D (image en niveaux de gris) et une image de la forme 3D (image de profondeur). Des descripteurs d'apparence et de profondeur sont extraits à partir de ces images. Afin de combiner les deux modalités, nous avons étudié trois stratégies de fusion. La fusion de descripteurs, la fusion de décisions et la fusion bi-niveaux. Trois scénarios sont donc possibles pour le processus de reconnaissance bimodale proposé. La méthode d'apprentissage utilisée pour la construction des classifieurs est la même pour les différents scénarios.

6.2 Extraction des descripteurs 2D et 3D

Dans le cadre de cette thèse, la description de l'apparence du visage est effectuée en utilisant une méthode de l'état de l'art. Les vecteurs caractéristiques des données 2D (images en niveaux de gris) sont obtenus en appliquant les LBP, dont l'efficacité a été montrée dans un grand nombre de travaux de l'état de l'art [67]. Par ailleurs, notre principale contribution concerne la description des données de profondeur. Elle consiste à proposer un nouveau descripteur, appelé *Depth Local Binary Pattern, DLBP*, dédié aux images de profondeur. Ce descripteur est inspiré du principe du LBP que nous avons étendu pour représenter les données de profondeur du visage. Contrairement aux méthodes de description géométriques basées sur le calcul des courbes et des caractéristiques 3D de la surface (nécessitant des données 3D de haute qualité), la méthode de description proposée tolère des données imprécises ou légèrement bruitées. Ainsi, elle n'impose que peu de contraintes concernant les équipements d'acquisition utilisés. Les données peuvent être obtenues par des scanners 3D, des caméras infrarouges ou des caméras stéréoscopiques.

6.3 Descripteur d'images de profondeur

Comme nous l'avons déjà évoqué dans l'état de l'art, le descripteur LBP a été utilisé dans beaucoup de travaux récents pour la reconnaissance 3D du visage [91, 145, 64, 133], en raison de son succès dans la reconnaissance 2D pour sa simplicité et son pouvoir discriminant. Lors de son application à la 3D, le LBP a donné des résultats satisfaisants quand il a été utilisé seul, et sa fusion avec d'autres descripteurs a donné des performances encore meilleures. Cependant, le descripteur LBP représente des formes 3D ayant des amplitudes différentes avec les mêmes motifs. Ceci est dû à sa méthode de codage qui n'utilise que les signes des différences de profondeur entre un pixel et son voisinage. Le fait de ne pas considérer les valeurs de ces différences conduit donc à représenter des formes ayant différentes magnitudes de la même façon. Une alternative, le 3DLBP, a été proposée [69] afin d'augmenter le pouvoir discriminant de LBP pour les formes 3D, comme indiqué dans la Section 3.5.3

de l'état de l'art. Cette extension a montré l'utilité de l'information de la magnitude dans la description des données 3D. Cependant, elle souffre d'inconvénients tels que la grande taille des vecteurs caractéristiques utilisés, la sensibilité du descripteur aux petites variations de profondeur et l'impossibilité de l'extension multi-échelles du descripteur. En effet, cette méthode n'est applicable que pour un rayon inférieur ou égal à 2, et par conséquent elle ne prend en considération que les variations locales de la profondeur du visage (cf. Section 3.5.3). Or, la variation locale de la profondeur est très faible, étant donné l'aspect lisse de la profondeur du visage, ce qui ne permet pas une bonne discrimination des visages.

Pour remédier aux inconvénients de ce descripteur, nous proposons une nouvelle extension du descripteur LBP conventionnel, le DLBP (*Depth Local Binary Pattern*), qui décrit la forme du visage en prenant en compte le signe et la magnitude de la différence de profondeur du voisinage. Ceci donne une description plus fine de la forme 3D et augmente considérablement son pouvoir discriminant. Par rapport aux autres extensions de LBP proposées pour la reconnaissance 3D, le DLBP offre, en plus de son efficacité, trois avantages principaux :

- l'exploitation de l'information de magnitude, tout en conservant le principe de simplicité du LBP conventionnel ;
- la stratégie multi-échelles est facilement applicable, afin de considérer les relations de voisinage à des échelles variables.
- la compacité de la représentation, la taille du descripteur est deux fois moindre que celle du 3DLBP.

6.3.1 Définition du DLBP

Le principe de ce descripteur repose sur deux étapes : le codage du signe et de la magnitude. Un pixel $p(x,y)$ est donc représenté par deux codes c^s et c^m. Afin de coder le signe du voisinage, nous utilisons le même principe que le LBP conventionnel. c^s est obtenu en parcourant les pixels du voisinage et en générant des bits reflétant le signe de la différence de profondeur *DD* (*Depth Difference*) entre le pixel central et chacun de ses pixels voisins (voir la Figure 6.1).

Le codage de la magnitude consiste à générer une séquence binaire en comparant la valeur absolue de la magnitude de la profondeur entre le pixel et son voisinage à un certain seuil S^m. Cette suite est ensuite lue de gauche à droite afin de construire le code décimal représentant la magnitude c^m. La Figure 6.1 montre un exemple de calcul de ce descripteur pour un pixel d'une image de profondeur avec un voisinage $V = 8$, un rayon $R = 1$ et un seuil de magnitude $S^m = 3$.

Comme montre la Figure 6.1, les magnitudes du voisinage sont codées dans le même ordre que les signes. Deux bits de même position dans les deux séquences binaires, représentant le signe et la magnitude, sont attribués à chaque pixel. Ceci permet une représentation cohérente et compacte des deux informations pour l'ensemble du voisinage. La Figure 6.2 montre un exemple des deux matrices M_{c^s}, M_{c^m} (encodant respectivement le signe et la magnitude des pixels) obtenues en applicant $DBLP_{(1,8)}$ sur tous les pixels d'une carte de profondeur donnée.

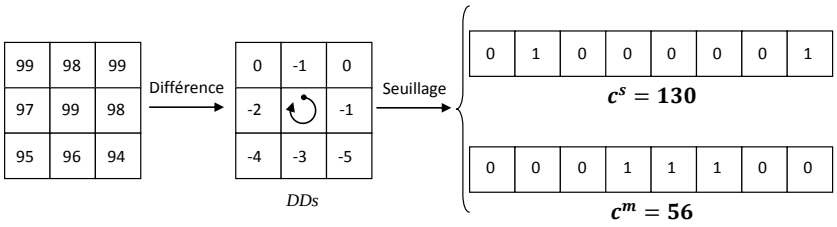

FIGURE 6.1: Illustration de l'extraction de c^s et c^m pour un pixel donné, avec $R = 1$, $V = 8$ et $S^m = 3$.

FIGURE 6.2: Exemple d'extraction de $DLBP_{(1,8)}$: (a) image de profondeur (b) matrice des codes de signes M_{c^s} (c) matrice des codes de magnitudes M_{c^m}.

6.3.2 Stratégie multi-échelles

En observant les images de profondeur des visages, il est aisé de constater que ces surfaces lisses et continues ne présentent pas de forts contrastes locaux en profondeur. Ainsi, l'utilisation de l'information locale peut être insuffisante pour extraire des données discriminantes des visages. Il se révèle donc utile de considérer le voisinage situé à une distance suffisamment grande afin que le contraste de profondeur soit plus important. Pour ceci, nous proposons d'étendre notre descripteur à des rayons plus larges en utilisant une stratégie multi-échelles :

$$DLBP_{(R,V)}(p) = \begin{pmatrix} c^s_{(R,V)}(p) \\ c^m_{(R,V)}(p) \end{pmatrix} = \begin{pmatrix} \sum_{i=0}^{V-1} s(p_i - p)2^i,\ s(k) = \begin{cases} 1 & \text{si} \quad k \geq 0 \\ 0 & \text{sinon} \end{cases} \\ \sum_{i=0}^{V-1} m(|p_i - p|)2^i,\ m(k) = \begin{cases} 1 & \text{si} \quad k \geq S^m_{(R,V)} \\ 0 & \text{sinon} \end{cases} \end{pmatrix} \tag{6.1}$$

où :

- p_i est le $i^{\text{ième}}$ pixel voisin dont la position par rapport à p est définie selon R et V (c.f. Equation 3.2) ;
- $S^m_{(R,V)}$ est le seuil de magnitude.

6.3.3 Calcul du seuil

Le valeur $S^m_{(R,V)}$ utilisée pour le seuillage de la magnitude est calculée automatiquement afin d'exploiter l'information de profondeur à différentes échelles. Pour ceci, nous appliquons une analyse statistique sur les données 3D afin de calculer la valeur optimale du seuil de magnitude $S^m_{(R,V)}$. Cette analyse permet de maximiser le pouvoir de discrimination du descripteur tout en négligeant le bruit éventuellement présent dans les cartes de profondeur.

Afin de trouver le seuil optimal de la magnitude, la première étape consiste à calculer un ensemble de gradients à différentes échelles et suivant différentes directions en fonction des paramètres (R,V) de DLBP. Nous utilisons la notation MSD-Gradient$_{(R,V)}$ (*gradient multi-échelles et multi-directions*) pour désigner les images de gradient calculées sur une image de profondeur. Le MSD-Gradient$_{(R,V)}$ d'un rayon R et voisinage V pour un pixel donné p d'une carte de profondeur est calculé comme :

$$\text{MSD-Gradient}_{(R,V)}(p) = \frac{1}{V}\sum_{i=0}^{V-1}|p-p_i| \tag{6.2}$$

Enfin, nous calculons la médiane de l'ensemble des valeurs MSD-Gradient$_{(R,V)}$ non nulles calculées sur l'ensemble des cartes de profondeur. Celle-ci est donc utilisée comme un seuil de magnitude $S^m_{(R,V)}$ pour le codage de $DLBP_{(R,V)}$.

La méthode proposée pour le calcul automatique du seuil de magnitude permet un codage adaptatif de la magnitude du voisinage à différentes échelles. Elle garantit également l'utilisation des données importantes uniquement, en ignorant les valeurs bruitées présentes dans la carte de profondeur grâce à l'utilisation de la médiane comme un seuil du codage. Nous ne prenons donc pas en compte la valeur de la différence de profondeur que le pixel forme avec son voisinage, mais l'intervalle où se trouve cette valeur – au-dessus au au-dessous du seuil. Un autre avantage repose dans l'invariance du descripteur proposé à la résolution des cartes de profondeur. Par conséquent, il peut être appliqué à des cartes de basse résolution obtenues à partir d'équipements moins précis que les scanners 3D comme les capteurs stéréoscopiques ou Kinect de Microsoft.

6.3.4 Construction des histogrammes

Afin que le descripteur soit moins sensible aux translations du visage, nous suivons la même stratégie utilisée pour l'application du LBP sur les images 2D du visage [5] qui consiste à exprimer le LBP sous la forme d'histogrammes locaux. L'utilisation des histogrammes permet aussi de diminuer la taille du descripteur et par conséquent le temps de calcul inhérent à la classification. Nous découpons donc les deux matrices M_{c^s} et M_{c^m} en un certain nombre de régions sur lesquelles nous calculons des histogrammes élémentaires (c.f. Equation 6.3) qui sont par la suite concaténés afin de créer le vecteur DLBP final. Dans la Figure 6.3 nous illustrons l'extraction des histogrammes à partir des deux matrices M_{c^s} et M_{c^m}.

$$h(b) = < n_b >, b = 1,...,B \tag{6.3}$$

où :
– B est le nombre de partitions de l'histogramme (nombre de codes DLBP) ;
– n_b est le nombre d'observations du code b.

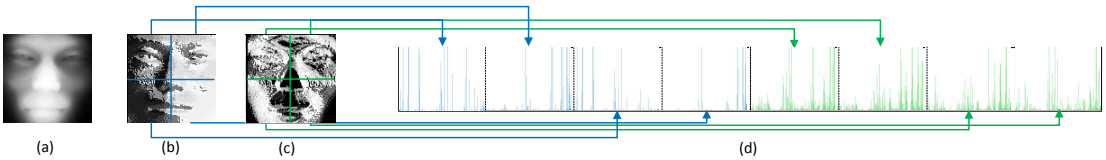

FIGURE 6.3: Exemple d'extraction d'histogrammes à partir des deux matrices de signes et de magnitudes : (a) image de profondeur (b) matrice des signes M_{c^s} (c) matrice des magnitudes M_{c^m} (d) histogramme extrait.

Le descripteur proposé permet une meilleure discrimination entre les formes 3D dont les topologies se ressemblent grâce à la prise en considération de la magnitude en plus du signe du voisinage. La stratégie multi-échelles permet de faire ressortir le contraste en profondeur de la forme 3D du visage qui n'est pas détecté localement. Enfin, le descripteur conserve la simplicité du LBP conventionnel.

6.4 Prédiction de l'identité par fusion

Nous décrivons dans cette section le processus proposé pour la combinaison des deux modalités 2D et 3D pour la reconnaissance de visages. Dans le cadre de notre travail, trois stratégies de fusion sont envisagées. La fusion de descripteurs, la fusion de décisions et la fusion bi-niveaux. L'étape de classification peut être effectuée par une méthode d'apprentissage quelconque retenue pour les trois stratégies.

6.4.1 Fusion de descripteurs

La première stratégie que nous avons adoptée consiste en la fusion de descripteurs. L'extraction de descripteurs à partir des images d'apparence (niveau de gris) et de profondeur d'un visage dont l'identité est inconnue est d'abord effectuée. Les deux vecteurs caractéristiques LBP et DLBP obtenus sont ensuite concaténés afin d'en construire un seul. Un classifieur est obtenu en effectuant une étape d'apprentissage sur les descripteurs fusionnés. Ce classifieur est utilisé pour l'identification d'un visage de test. La Figure 6.4 illustre les étapes de ce premier scénario.

6.4.2 Fusion de décisions

Dans la deuxième stratégie, deux processus de classification sont effectués indépendamment pour les deux ensembles de descripteurs LBP et DLBP. Nous construisons donc un classifieur par modalité. Lors de l'identification, les décisions obtenues par les deux classifieurs sont fusionnées afin d'en déduire l'identité du visage (voir la Figure 6.5). Étant donné

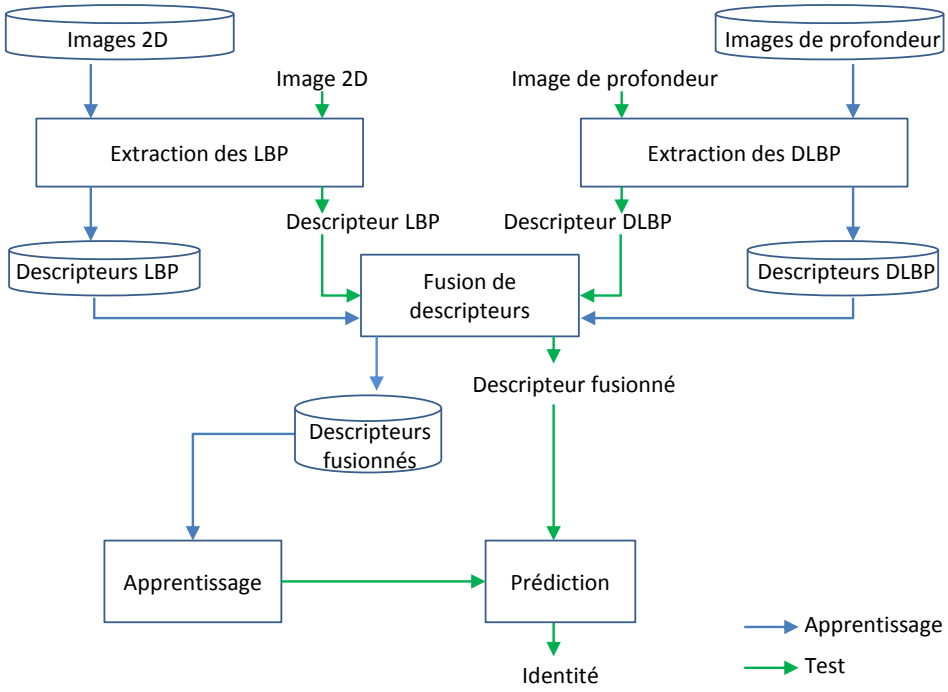

FIGURE 6.4: Fusion de descripteurs.

un visage représenté par deux images d'apparence et de profondeur (I_1 et I_2), les deux classifieurs M_1 et M_2 obtenus par l'étape d'apprentissage fournissent respectivement deux décisions $M_1(I_1)$ et $M_2(I_2)$. Afin de calculer la décision finale, nous proposons d'utiliser une méthode basée sur le principe de vote majoritaire pondéré [157]. Cette dernière est la méthode de fusion la plus connue et la plus simple à mettre en œuvre. Notons $M_j(I_j) = i$ le fait que le classifieur M_j attribue la classe $i (i \in 1,2,...,n)$ à l'image I_j, où $j \in 1,2,...,m$ et m est le nombre de modalités (égale à 2 dans le cas de fusion bimodale). Une fonction indicatrice F est associée à chaque classifieur et est définie par l'équation suivante :

$$F_i^j(I_j) = \begin{cases} 1 & si \quad M_j(I_j) = i, \\ 0 & sinon. \end{cases} \tag{6.4}$$

La combinaison des deux classifieurs s'écrit donc :

$$F_i^E = \sum_{j=1}^{m} \alpha_{ij} F_i^j(I_j) \tag{6.5}$$

pour tout $i \in 1,2,...,n$. Les poids α_{ij} représentent la fiabilité du classifieur j pour une décision donnée (classe i). Ces poids constituent les taux de reconnaissance de chacun des classifieurs obtenus pour chaque classe lors de l'étape de l'apprentissage. La décision finale (l'identité) consiste à trouver $argmax_i(F_i^E)$.

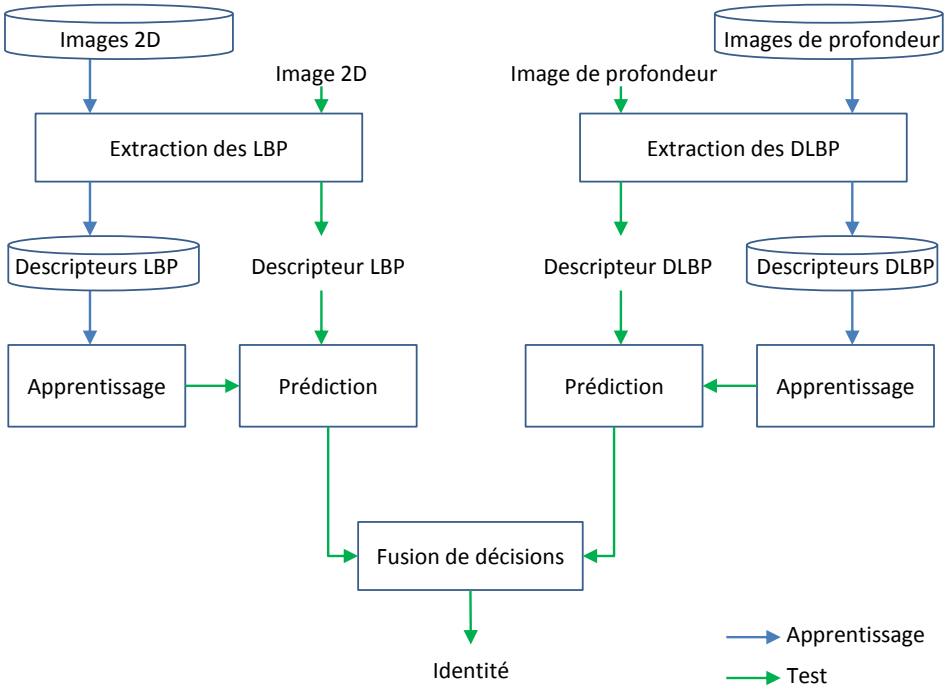

FIGURE 6.5: Fusion de décisions.

6.4.3 Fusion bi-niveaux

Le troisième scénario que nous proposons consiste à effectuer une fusion bi-niveaux en combinant les deux stratégies de fusion afin de bénéficier de leurs avantages respectifs (voir la Figure 6.6). Pour un visage donné, trois décisions sont donc obtenues à partir du classifieur 2D, du classifieur de profondeur et du classifieur des descripteurs fusionnés. Ces trois décisions sont fusionnées en utilisant la méthode de vote majoritaire décrite ci-dessus.

6.5 Conclusion

Dans ce chapitre, nous avons présenté l'approche proposée pour la reconnaissance de visages. L'approche utilise les deux modalités 2D et 3D afin d'améliorer la performance des systèmes de reconnaissance monomodale. Le descripteur LBP dont l'efficacité a été prouvée dans les travaux de l'état de l'art a été utilisé pour la description des images 2D. Un nouveau descripteur, appelé *DLBP*, est proposé afin d'extraire les vecteurs de caractéristiques discriminants à partir des images de profondeur. Le DLBP intègre l'information de la magnitude en plus du signe des différences de profondeur du voisinage d'une manière compacte et cohérente. L'avantage principal de notre descripteur est qu'il est basé sur un processus de seuillage adaptatif de l'information de magnitude, ce qui permet de l'appliquer à différentes échelles. En effet, nous partons de l'hypothèse qu'à grande échelle, la variation

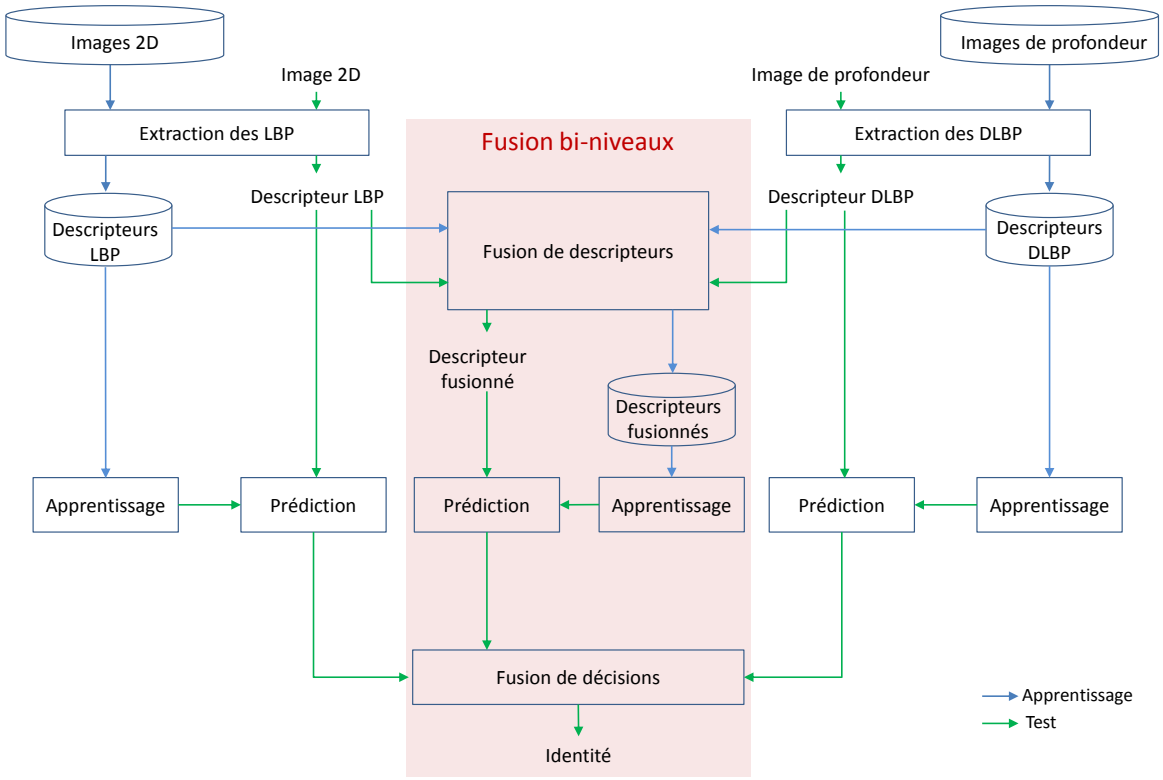

FIGURE 6.6: Fusion bi-niveaux.

de la profondeur est plus discriminante, étant donné l'aspect lisse de la surface 3D du visage. La reconnaissance de visages est effectuée par une stratégie de fusion bi-niveaux. Celle-ci permet de prendre en considération la relation entre les deux modalités en deux étapes (avant et après l'apprentissage).

Dans le travail présenté dans ce chapitre, nous considérons que l'étape d'acquisition est déjà effectuée et que les données 3D sont disponibles. Ces données peuvent être acquises par tout type de capteurs 3D (scanner 3D, Kinect, caméra stéréoscopique, etc.). Ceci n'influence pas le mode de fonctionnement de notre approche de reconnaissance puisqu'aucune hypothèse n'est faite sur la qualité des données 3D.

Dans la partie suivante, nous présentons la procédure expérimentale suivie afin d'évaluer les différentes contributions proposées dans le cadre de cette thèse, au niveau de l'acquisition (reconstruction stéréoscopique de visage) et au niveau de la reconnaissance (extraction de descripteurs et fusion).

Troisième partie

Expérimentations, résultats et discussion

Chapitre 7

Élaboration d'une collection de tests

Sommaire

7.1 Contexte et besoins 84
7.2 Matériel utilisé 84
7.3 Méthodologie 86
7.4 Annotation 87
7.5 Discussion et conclusion 88

7.1 Contexte et besoins

Avec le succès des méthodes 3D pour la reconnaissance de visages, beaucoup de collections de tests 3D publiques ont été créées afin d'évaluer les différentes méthodes 3D proposées. Ces collections sont réalisées via une acquisition active à base d'équipements onéreux comme les scanners 3D. Ces équipements permettent d'obtenir des données d'une grande qualité et des modèles 3D complets. Ces collections permettent de tester en conditions idéales les approches développées. Cependant, ces conditions sont relativement contraintes du fait de la spécificité du capteur (scanner 3D onéreux et donc peu répandus), et des conditions d'acquisition (temps de numérisation et nécessité de la collaboration de la personne à identifier). Ainsi, ces collections ne permettent pas de tester les approches dans des conditions d'applications réelles (dispositifs de capture rapide, peu onéreux, et donc souvent peu précis). Comme nous l'avons déjà vu dans la Section 2.2, les données 3D peuvent être obtenues à l'aide d'équipements moins contraignant comme les caméras stéréoscopiques, infrarouge (Kinect) ou de type temps-de-vol (*Time-f-Flight, ToF*). Certes, la qualité des données obtenues avec ce genre d'équipements est moins bonne que celle obtenue avec des scanners 3D. De plus, l'information 3D est représentée par une carte de profondeur qui donne une représentation partielle du modèle 3D du visage. Néanmoins, leur utilisation permet d'obtenir une information de forme 3D plus rapidement et avec beaucoup moins de contraintes, ce qui permet d'élargir le champ d'applications de la reconnaissance 3D de visages ou bimodale. C'est pour cela que l'intérêt des recherches récentes se focalise de plus en plus sur l'utilisation de ces équipements pour la reconnaissance de visages. Cependant, il existe très peu de collections de données 3D construites à l'aide de ces équipements.

Cela nous a conduits à développer une nouvelle collection de tests en utilisant ces équipements peu contraints. Ces équipements ont été utilisés pour l'acquisition de données visuelles et de profondeur de 64 personnes, en faisant varier les conditions d'éclairage, les expressions faciales, ainsi que la pose de la tête. Les données ont été annotées manuellement. La collection ainsi créée est riche en termes de variations intra-classes et inter-classes, et répond aux besoins du paradigme que nous proposons pour la reconnaissance faciale. Il est à noter qu'elle peut aussi être utilisée pour l'évaluation d'autres méthodes d'analyse de visages comme l'estimation de la pose et la reconnaissance d'expressions faciales.

7.2 Matériel utilisé

Les données de la collection élaborée sont obtenues à l'aide d'un système d'acquisition composé de trois capteurs différents (voir la Figure 7.1) :

1. **Caméra 3D infrarouge (Kinect)** : cet équipement, conçu par *Microsoft*, contient une caméra qui permet l'acquisition d'une image couleur de la scène ainsi qu'un capteur infrarouge offrant une image de profondeur de cette scène (voir la Section 2.2). Ce capteur est composé d'un émetteur de lumière infrarouge couplé à une caméra infrarouge à capteur CMOS (QVGA 320x240, 16 bits). L'émetteur de lumière infrarouge émet des rayons qui sont réfléchis par les objets de la scène. La caméra acquiert la

FIGURE 7.1: Système d'acquisition : (a) caméra infrarouge (Kinect) (b) caméra ToF (SR4000) (c) caméra stéréoscopique (Bumblebee XB3).

quantité de lumière infrarouge réfléchie par chaque point de la scène afin de construire la carte de profondeur, sous l'hypothèse que plus l'objet est loin, plus la quantité de rayonnement infrarouge réfléchi sera faible. Il est donc possible de cartographier la distance de tout objet éloigné de 1 à 4-5 mètres de la caméra.

2. **Caméra 3D *Time-of-Light (ToF)* (SR4000)** : il s'agit d'une caméra de type *ToF* conçue par la compagnie *Mesa Imaging*, qui permet de fournir en temps réel les informations 3D des objets de la scène (voir la Section 2.2). Son fonctionnement consiste à illuminer la scène et les objets par des impulsions lumineuses infrarouges, et à mesurer le temps que cette impulsion prend pour effectuer le trajet aller-retour entre la caméra infrarouge et l'objet. Le *temps de vol* de cette impulsion lumineuse est directement proportionnel à la distance entre la caméra et l'objet visé, ce qui permet d'obtenir une image de profondeur de la scène.
3. **Caméra stéréoscopique (Bumblebee XB3)** : il s'agit d'une caméra stéréoscopique conçue par la société *Point Grey*. Cette caméra a la particularité d'être *multi-baselines* [1] afin d'améliorer la flexibilité et la précision de l'acquisition. Elle dispose de 3 capteurs de 1,3 mégapixel. La grande *baseline* offre plus de précision à des distances plus grandes, tandis que la petite *baseline* améliore l'appariement de petits intervalles.

Trois sous-collections ont été créés par ce système d'acquisition : une collection stéréoscopique appelée FoxStereo et deux collections bimodales 2D-3D appelées FoxKinect et FoxTOF. Les données 3D sont sous forme de cartes de profondeur.

- **FoxStereo** : cet ensemble contient les images stéréoscopiques obtenues par le capteur Bumblebee XB3. Pour chaque personne, trois images (gauche, droite et milieu) de dimension 640×480 sont acquises simultanément. La Figure 7.2 montre un exemple des données acquises avec ce capteur.
- **FoxKinect** : l'ensemble des données de cette collection contient des images d'apparence 2D ainsi que les cartes de profondeur correspondantes. La dimension des images

1. Deux *baselines* de 6 cm et 12 cm sont disponibles.

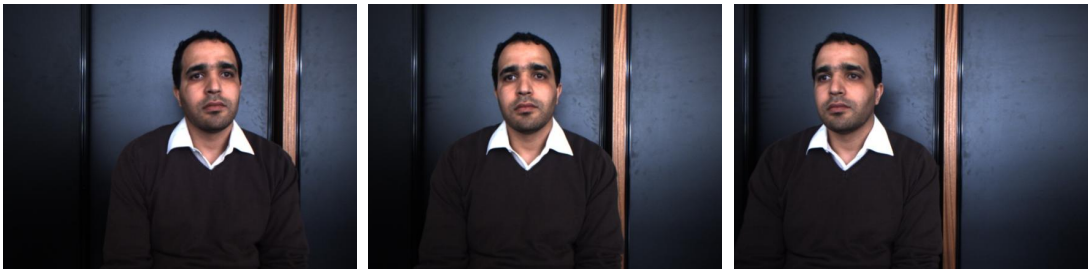

FIGURE 7.2: Exemple de triplet d'images acquises à l'aide du capteur stéréoscopique (Bumblebee XB3).

couleur est de 640×480, et 320×240 pour les images de profondeur. La Figure 7.3 montre un exemple des données obtenues par Kinect.

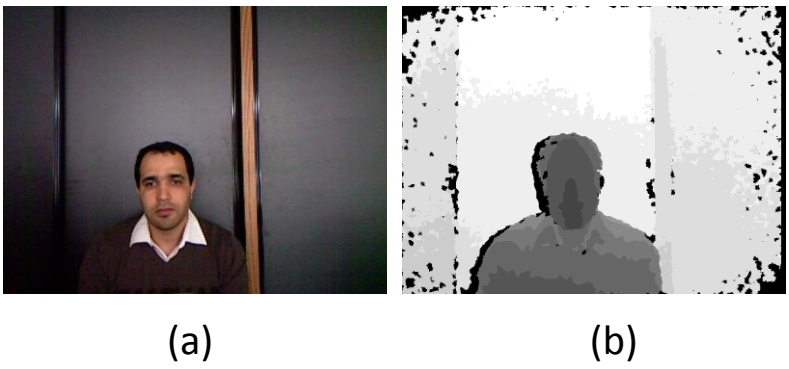

FIGURE 7.3: Exemple des images obtenues à l'aide de la caméra infrarouge (Kinect) : (a) image couleur (b) image de profondeur.

– **FoxTOF** : la caméra TOF fournit trois images par prise de vue : une image issue du capteur infrarouge, une image de profondeur et une matrice de confiance. Cette dernière associe une valeur de confiance à chaque pixel de l'image de profondeur. La dimension des trois images est de 176×144. Ce capteur est celui qui donne la résolution la plus basse des trois capteurs utilisés. La Figure 7.4 montre un exemple des données obtenues à l'aide de cette caméra.

7.3 Méthodologie

L'acquisition des données a eu lieu dans un environnement intérieur (bureaux du laboratoire LIFL). Au total, 64 personnes sont présentes dans la collection, dont 46 hommes (deux sont des jumeaux) et 18 femmes, âgées de 22 à 59 ans. Chaque personne est située à une

FIGURE 7.4: Exemple d'images obtenues à l'aide de la caméra ToF (SR4000) : (a) image infrarouge (b) image de profondeur (c) matrice de confiance (plus le pixel est clair, plus la confiance est bonne).

distance de 1 mètre du système d'acquisition [2]. Différentes variations en termes d'éclairage, d'expression et de pose sont effectuées afin de constituer les trois sous-collections décrites ci-dessus. Pour chaque personne, 40 prises de vues sont enregistrées. Elles correspondent à :

- 3 conditions d'éclairage : ambiant, frontal, latéral ;
- 7 expressions faciales : joie, tristesse, colère, dégoût, peur, surprise, *aucune* ;
- 30 poses résultant de la composition de neuf positions en *yaw* (de $-\frac{\pi}{2}$ à $\frac{\pi}{2}$, par tranches de $\frac{\pi}{8}$), selon trois positions en *pitch* (vers le bas, en face, vers le haut), plus deux positions en *roll* (gauche et droite).

Au total, la collection contient 2560 images. La Figure 7.5 montre un exemple des différentes variations acquises pour une personne donnée.

En plus des prises de vue statiques (images 2D et de profondeur), une prise de vue dynamique (séquence vidéo) est également enregistrée pour chaque personne, contenant toutes les variations (éclairage, expression et pose) pour chacune des trois collections de visages : FoxStereo, FoxKinect et FoxTOF.

7.4 Annotation

Une annotation manuelle de quatre points caractéristiques principaux (yeux, nez et bouche) a été effectuée afin de permettre aux utilisateurs de cette collection de localiser le visage, ainsi que ces points caractéristiques. La Figure 7.6 montre les quatre points annotés sur une image de visage. L'étiquette de chaque image encode les informations la concernant par six attributs :

- le numéro de la personne ;
- le numéro de l'image ;
- la pose ;
- l'éclairage ;

2. Il s'agit de la distance minimale pour permettre l'utilisation de Kinect.

FIGURE 7.5: Les différentes variations acquises pour chaque personne : (a) 3 pour l'éclairage (b) 7 pour l'expression (c) 30 pour la pose.

- l'expression faciale ;
- le type de l'image (image 2D ou de profondeur par exemple).

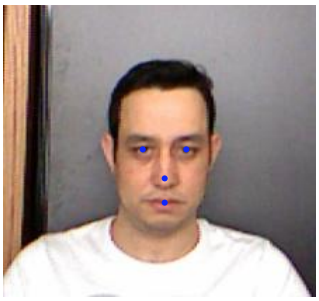

FIGURE 7.6: Annotation d'une image de la base de données.

7.5 Discussion et conclusion

L'information de profondeur s'est révélée être très utile dans le domaine de la vision par ordinateur et notamment dans les applications d'analyse de visages. Dans ce chapitre, nous

avons présenté une collection de test construite dans le cadre de cette thèse. La collection élaborée comprend des enregistrements statiques (images 2D et de profondeur) et dynamiques (séquences vidéos) de visages acquis sous différentes conditions de pose, d'éclairage et d'expression faciale. Les dispositifs utilisés permettent l'acquisition de l'information de profondeur avec moins de contraintes en termes de coût et de temps d'acquisition que les scanners 3D. L'ensemble des données de cette collection permet l'évaluation de différents types d'algorithmes d'analyse de visages grâce à la variété des données acquises notamment :

- reconstruction 3D du visage : les images stéréoscopiques peuvent être utilisées afin de reconstruire des cartes de profondeur de visages en utilisant différents algorithmes de reconstruction stéréoscopique. Une autre utilisation consiste à combiner les différentes cartes de profondeurs disponibles dans la collection afin d'avoir un modèle 3D complet ;
- reconnaissance 2D : comme la collection élaborée contient des images de couleur, elle peut être utilisée pour la reconnaissance 2D classique ;
- reconnaissance 3D, ou bimodale : en passant par une étape de reconstruction ou bien en utilisant directement les cartes de profondeur disponibles dans FoxKinect et FoxTOF, les cartes de profondeur peuvent être utilisées pour les algorithmes de reconnaissance 3D ou bimodale en les combinant avec les images couleurs ;
- reconnaissance d'expressions faciales et estimation de la pose : les méthodes d'estimation de la pose ou de reconnaissance d'expressions faciales, que ce soit en 2D, 3D ou 2D-3D, peuvent être évaluées en utilisant les différentes données de notre collection ;
- suivi et reconnaissance dynamique de visages : les vidéos disponibles dans notre collection peuvent être utilisées pour l'évaluation des algorithmes de suivi ou de reconnaissance dynamique de visages en utilisant les séquences couleur ou profondeur, ou en combinant ces deux modalités.

Dans la suite de cette partie, nous présentons l'évaluation de la méthode de reconstruction stéréoscopique proposée dans le chapitre 8. Dans le chapitre 9, nous nous intéressons à la partie reconnaissance où la méthode de reconnaissance bimodale est évaluée.

Chapitre 8

Évaluation de la méthode de reconstruction stéréoscopique de visages

Sommaire

8.1 Introduction 92
8.2 Détection du bruit 92
8.3 Estimation de la profondeur 94
8.3.1 Illustration des résultats de la méthode proposée 95
8.3.2 Comparaison aux méthodes stéréoscopiques 95
8.3.3 Comparaison aux autres méthodes de reconstruction 100
8.4 Conclusion 105

8.1 Introduction

Dans ce chapitre, nous évaluons qualitativement et quantitativement les résultats de la méthode de reconstruction de visages proposée. Les visages sont reconstruits selon la méthode d'estimation proposée décrite à la Section 5.2, et l'étape de débruitage proposée (voir la Section 5.3) est appliquée. Nous commençons par la présentation de l'évaluation de notre méthode de débruitage. Ensuite, l'évaluation de la qualité de reconstruction est effectuée sur différentes collections de tests.

8.2 Détection du bruit

La méthode de détection de bruit est évaluée dans cette section afin de montrer comment cette étape contribue à débruiter la carte de profondeur tout en préservant les données valides. Afin d'évaluer la classification des zones de l'image en zones bruitées ou non-bruitées que fournit l'algorithme de détection de bruit, nous avons calculé la matrice de confusion (voir le Tableau 8.1) sur un ensemble de 100 lignes de profondeur extraites à partir de quelques cartes de profondeur sur lesquelles nous avons généré des zones bruitées de différentes tailles afin de constituer une vérité terrain. Le nombre total de segments obtenus par l'étape de segmentation est 269, dont 108 sont bruités et 161 sont non-bruités.

Vérité terrain\Détection	Bruité	Non-bruité
Bruité	102	4
Non-bruité	6	157

TABLE 8.1: Matrice de confusion de la classification des zones de l'image en zones bruitées et non bruitées.

La matrice de confusion montre la capacité de l'algorithme à détecter les segments bruités d'une carte de profondeur donnée. La précision est de $96,28\%$. Nous notons que les classifications erronées données par la méthode de détection de bruit (6 faux positifs et 4 faux négatifs) correspondent aux cas où la partie bruitée est plus grande que la partie non-bruitée dans la ligne de profondeur. En effet, dans ce cas, les valeurs statistiques calculées sur la ligne de profondeur sont plus influencées par le bruit que par les données correctes et donc la classification est inversée.

Afin de montrer comment l'algorithme proposé permet d'identifier le bruit dans la carte de profondeur, nous comparons dans la Figure 8.1 le résultat du processus de débruitage de notre algorithme à deux méthodes basées sur le filtre médian : une méthode basée sur l'application du filtre médian sur la totalité de la carte de profondeur [15], et une méthode basée sur un débruitage local par le filtre médian (identification et suppression du bruit) [98]. Dans les deux cas, le filtre est basé sur un noyau k variable.

Nous pouvons voir que les résultats du débruitage en appliquant le filtre médian sur la totalité de la carte de profondeur semblent visuellement satisfaisants (Figure 8.1 (a)). Cependant, il n'est pas robuste à la taille du bruit. En effet, pour un noyau $k = 7$, seules les petites

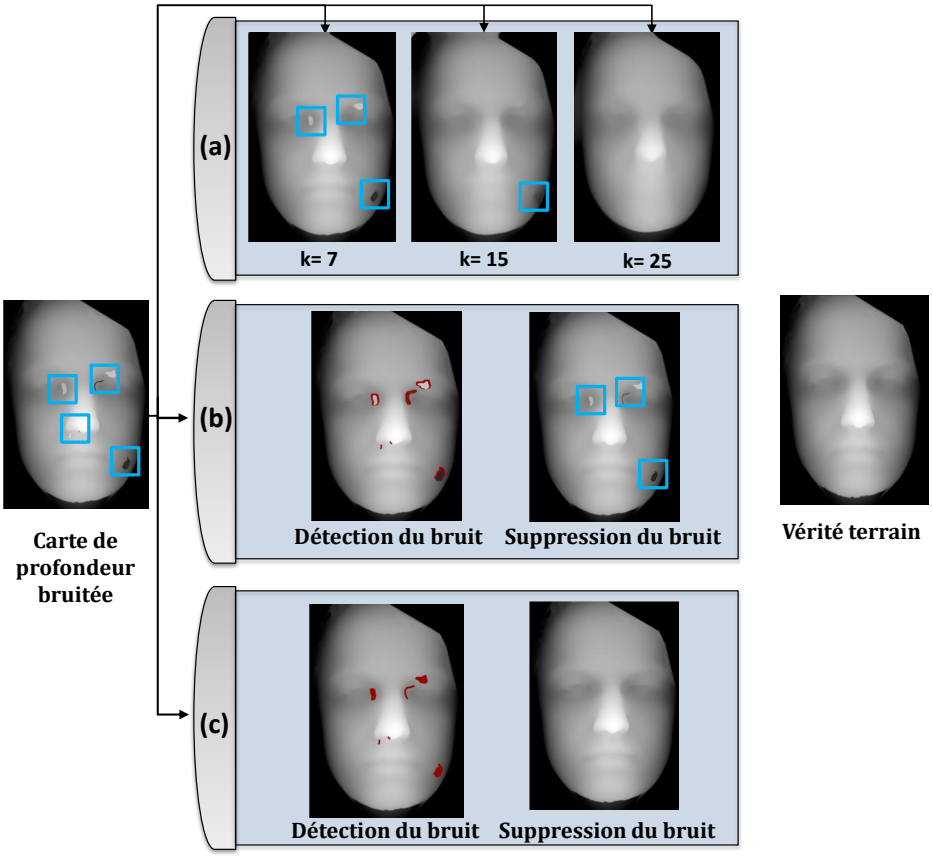

FIGURE 8.1: Débruitage de la carte de profondeur : (a) débruitage global par le filtre médian [15], (b) débruitage local par le filtre médian [98], (c) méthode proposée.

zones bruitées sont éliminées. Un noyau supérieur ($k = 25$) est nécessaire pour supprimer les grandes zones bruitées. De plus, l'image entière est traitée et les valeurs correctes de profondeur du reste des pixels sont modifiées. Par ailleurs, le débruitage local (Figure 8.1 (b)) corrige localement la carte de profondeur (parties bruitées) et préserve ainsi l'information de profondeur et les détails. Les bruits de petite taille (pics) sont détectés avec précision. Au niveau des zones bruitées de grande taille, seuls les bords sont identifiés comme du bruit. En effet, les données du voisinage pour le reste des pixels (loins des bords) ne permettent pas de détecter le bruit du fait qu'elles sont stables dans la région bruitée. Notre algorithme (Figure 8.1 (c)) est capable de détecter les zones bruitées de différentes tailles de la même manière car il est fondé sur des statistiques calculées, localement et globalement, sur la carte de profondeur estimée.

Afin de montrer comment l'étape de détection de bruit contribue à la préservation des données lors du débruitage, nous calculons l'erreur moyenne quadratique (*Root Mean Square Error, RMS*) [119] entre les cartes de profondeur et la carte de vérité terrain avant et après l'étape de débruitage (ce calcul est effectué sur les images illustrées dans la Figure 8.1). Ceci permet de montrer la variation de la RMS avant et après le débruitage. La RMS est calculée

selon l'équation suivante :

$$RMS = \left(\frac{1}{n} \times \sum_{x,y} (d_E(x,y) - d_T(x,y))^2 \right)^{\frac{1}{2}} \tag{8.1}$$

où :
- n est le nombre de pixels de la carte de profondeur ;
- $d_E(x,y)$ et $d_T(x,y)$ sont respectivement la valeur de profondeur estimée du pixel et celle de la vérité terrain.

Les résultats sont illustrés dans la Figure 8.2. Nous pouvons voir dans cette figure que l'erreur RMS obtenue sur les cartes de profondeur débruitées globalement est plus grande que celle obtenue par la carte bruitée. Ceci est dû au processus global utilisé par cette méthode qui affecte toute la carte de profondeur. Ce processus provoque ainsi la perte des données correctes de profondeur. Le débruitage local donne une RMS plus petite du fait qu'il est basé sur une étape de détection du bruit. L'erreur la plus petite est obtenue avec notre méthode. Ceci est dû à la précision du processus de détection du bruit proposé.

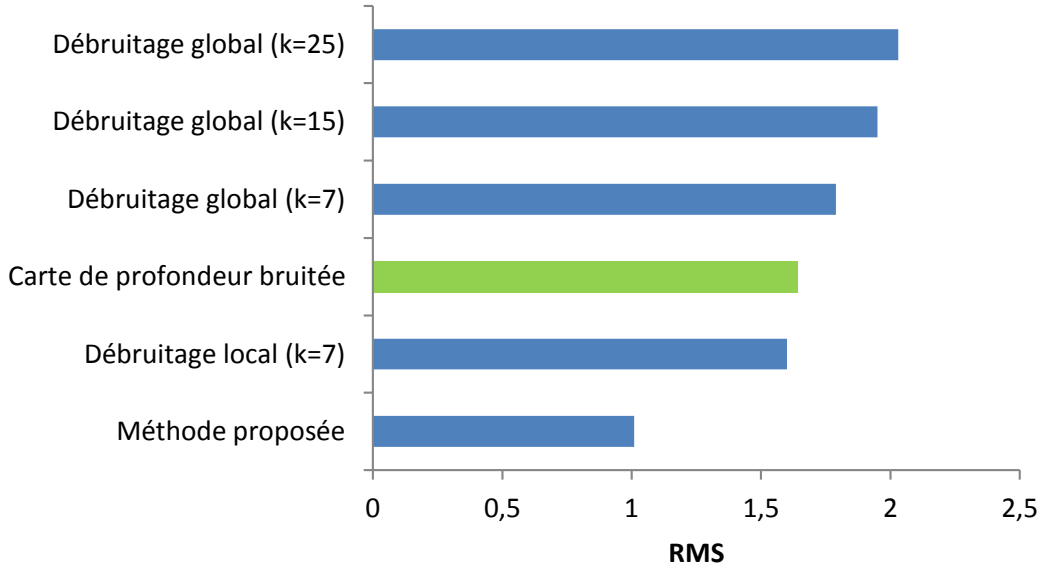

FIGURE 8.2: RMS entre les cartes de profondeur débruitées et la vérité terrain.

8.3 Estimation de la profondeur

Dans cette section, nous évaluons la méthode de reconstruction proposée sur différentes collections de test. Dans toutes les expérimentations réalisées, le processus suivant est appliqué. D'abord, les cartes de disparité sont estimées à partir des images stéréoscopiques.

Ensuite, les cartes de profondeur sont calculées. Enfin, l'étape de post-traitement consistant à détecter et supprimer le bruit dans les cartes de profondeur estimées est appliquée en utilisant l'algorithme de post-traitement proposé. L'évaluation est effectuée en trois étapes :

- nous commençons par une illustration des résultats obtenus sur des exemples de paires stéréoscopiques obtenues à l'aide du capteur Bumblebee XB3 (cf. Section 7.2) ;
- ensuite, notre méthode (spécifique aux visages) est comparée à deux méthodes d'appariement stéréoscopique génériques (destinées aux objets en général) notamment la méthode locale d'appariement de blocs (*block matching, BM*) [20] et la méthode globale de coupe de graphe (*graph cut, GC*) [80]. La première (BM) est la méthode sur laquelle notre approche est basée. Ceci nous permet de montrer comment l'intégration des a priori obtenus par l'étape de la construction du modèle de disparité améliore le processus d'estimation de la profondeur sans l'utilisation d'un modèle déformable 3D ni d'autres processus exigeant un temps de traitement supplémentaire. La deuxième (GC) est utilisée dans la comparaison comme un exemple de méthode globale ;
- enfin, nos résultats sont comparés à ceux obtenus par d'autres méthodes de reconstruction de visages, basées notamment sur les modèles 3D comme dans les travaux de Sun *et al.* [127, 128], Koo *et al.* [81] et Fortuna *et al.* [50]. Une évaluation quantitative sur la collection Bosphorus [118] est donnée par ces derniers. Ainsi, nous comparons la précision de notre méthode de reconstruction à ces travaux, en utilisant la même configuration.

8.3.1 Illustration des résultats de la méthode proposée

Dans la Figure 8.3, nous montrons les résultats obtenus à l'aide de notre méthode sur quelques exemples de visages [1]. La taille du visage est de 120×180 pixels, et la taille du bloc d'appariement est de 11×11. L'intervalle de disparité est défini automatiquement en fonction du modèle de disparité du visage construit à partir de l'ASM. Un masque elliptique est appliqué automatiquement pour supprimer l'arrière-plan des cartes illustrées. La figure montre l'étape d'ajustement de l'ASM et la carte de disparité reconstruite avec notre méthode pour ces visages. Nous pouvons voir que le processus d'estimation de disparité donne un résultat visuellement bon et peu bruité pour la vue frontale. En cas de petites variations de pose (moins de 20°), notre méthode donne toujours une bonne estimation puisque le processus d'ajustement de l'ASM est toujours possible. Toutefois, lorsque la variation de pose est grande, l'ASM ne peut pas trouver tous les points nécessaires pour son ajustement sur le visage et, par conséquent, le modèle de disparité ne peut pas être construit. Dans ce cas, la carte de disparité ne peut pas être calculée par notre méthode.

8.3.2 Comparaison aux méthodes stéréoscopiques

Dans cette section, nous commençons par une comparaison qualitative effectuée sur des exemples de la collection de tests *3D Texas database* [56]. Cette collection contient un total

1. Les images proviennent du capteur Bumblebee XB3

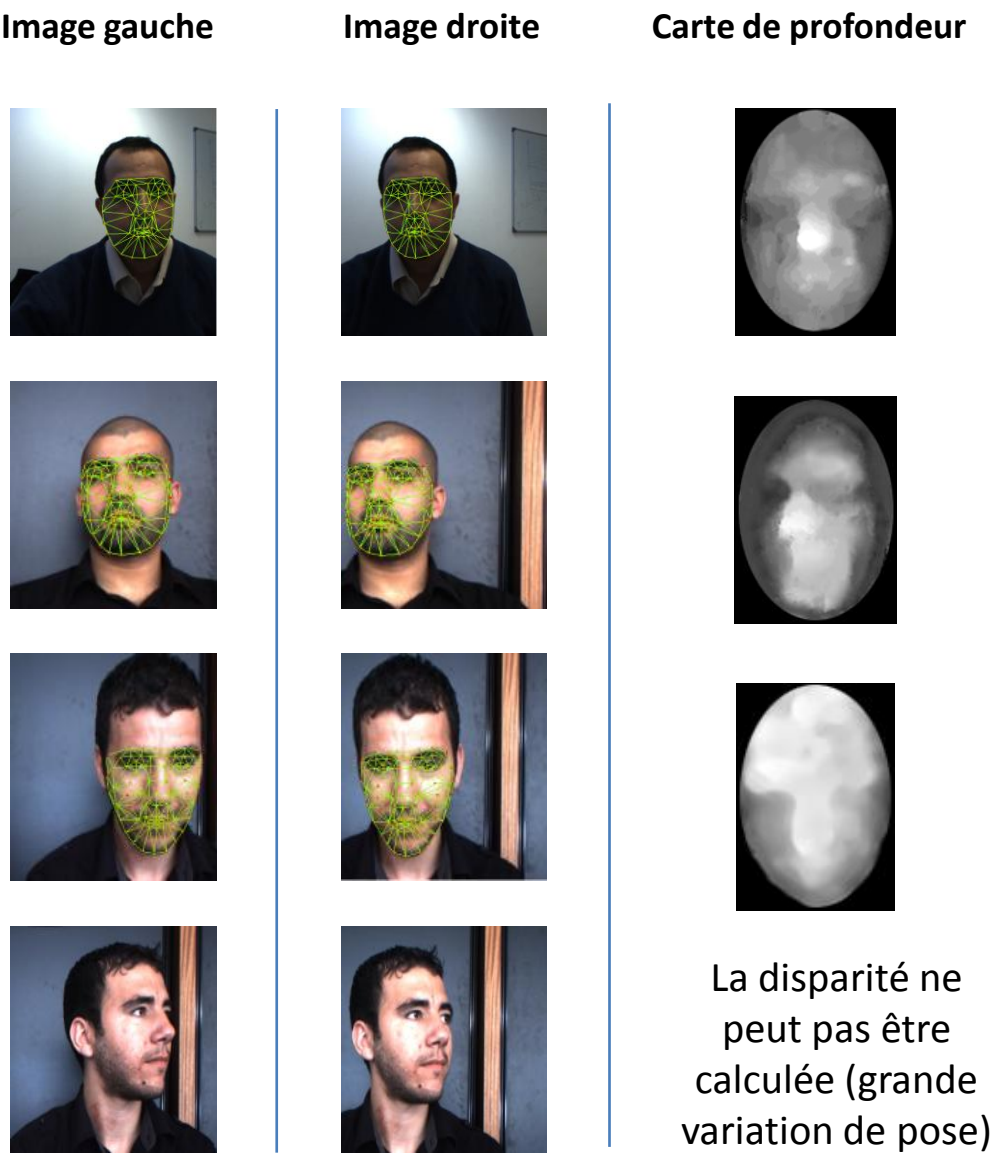

FIGURE 8.3: Cartes de disparité pour des visages sous petites variation de pose.

de 1149 images couleur de 118 personnes, ainsi que les images de profondeur correspondantes. Toutes les prises de vue sont de face. Pour chaque personne, la variabilité des images provient principalement de la variation d'éclairage, et dans une moindre mesure, de quelques variations des expressions faciales. L'avantage de cette collection est qu'elle contient les images de profondeur de vérité terrain qui peuvent être utilisées comme référence lors de la comparaison. Ensuite, différentes expérimentations sont réalisées en vue de donner une évaluation quantitative de la méthode proposée.

Afin d'effectuer nos expérimentations, les paires stéréoscopiques sont synthétisées à par-

tir des données de la vérité terrain fournies dans cette collection. Ceci est effectué en construisant un maillage 3D texturé à partir des images de profondeur et couleur et en le projetant sur deux plans stéréoscopiques, en utilisant des outils infographiques [2]. Les images sont de taille 501×751 pixels. Les cartes de disparités sont ensuite calculées par les trois méthodes stéréoscopiques (BM, GC et notre méthode). La taille du bloc d'appariement (utilisée par notre méthode et par la méthode BM) de 11×11 est conservée. Comme les méthodes BM et GC sont liées à l'intervalle de disparité, nous avons réalisé des tests expérimentaux avec des valeurs différentes et nous avons choisi l'intervalle qui donne les meilleurs résultats.

Comparaison qualitative sur des paires stéréoscopiques de Texas

Dans la Figure 8.4, nous montrons les cartes de profondeur obtenues à l'aide des trois méthodes ainsi que les cartes de vérité terrain sur deux exemples de visages. Selon cette

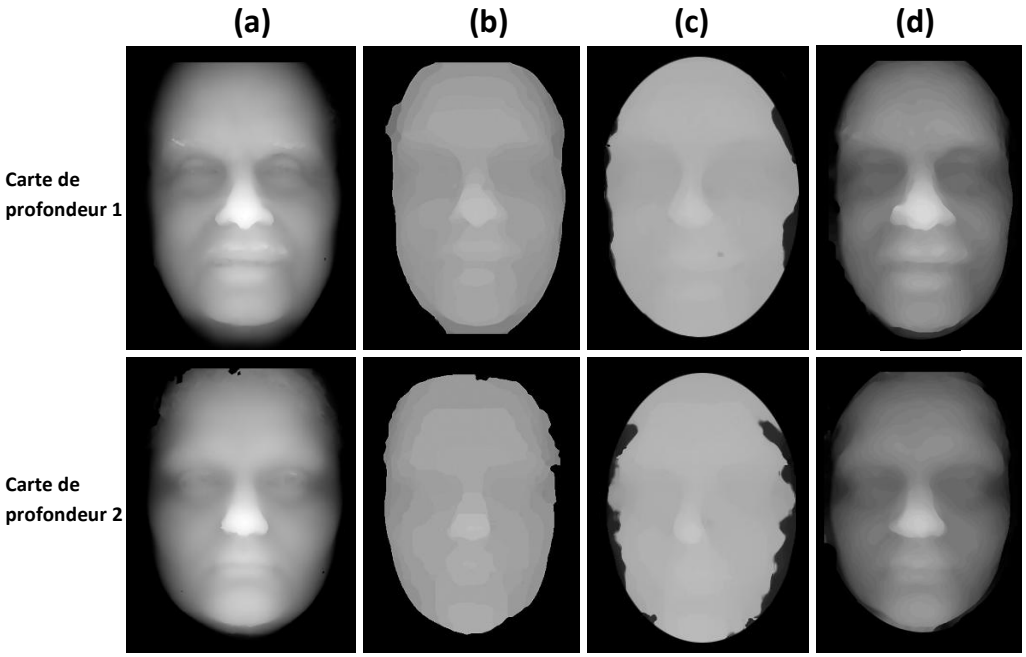

FIGURE 8.4: Cartes de profondeur : (a) vérité terrain (b) *graph cut* (c) *block matching* (d) notre méthode.

figure, nous pouvons voir que visuellement notre méthode donne de meilleurs résultats que la méthode BM en termes de régularité et de précision. Ceci montre que le modèle de disparité construit dans notre méthode a permis d'améliorer le processus d'appariement. De plus, elle nécessite moins de temps de traitement, puisque nous associons, pour chaque point, un petit intervalle de disparité défini en fonction de la partie topologique du visage à laquelle le point appartient.

2. Nous avons utilisé la bibliothèque OpenGL pour effectuer cette partie du travail.

La Figure 8.4 illustre aussi les résultats obtenus à l'aide de la méthode globale, GC [80], qui est réputée donner de très bons résultats d'estimation [119]. Cette méthode semblent donner de meilleurs résultats que la méthode BM. Cependant, nous pouvons voir que la profondeur de visage est segmentée et donne moins de détails que notre méthode sur les zones topologiques d'intérêt (nez, yeux, etc.). La méthode de coupe de graphe représente l'image sous forme de graphe et cherche d'une façon itérative des coupes dans ce graphe qui correspondent à différentes valeurs de disparités. Lorsque l'image est constituée d'une scène contenant des objets différents, la méthode donne une bonne estimation. Toutefois, étant donné que le visage est une surface continue et lisse, les résultats de cette méthode contiennent des zones planes avec des coupures brusques, car ils proviennent d'un processus de segmentation. Cela conduit à la perte d'information de profondeur et de l'aspect trop lisse de la carte de disparité. Par ailleurs, ce n'est pas le cas dans nos résultats, qui sont les plus proches de la vérité terrain.

Comparaison quantitative sur toute la collection Texas

Afin de comparer quantitativement les résultats obtenus sur la collection Texas, nous calculons l'erreur moyenne quadratique (*root mean square error, RMS*, cf. Équation 8.1) et le pourcentage de faux appariements (*percentage of bad matching pixels, PBM*, cf. Équation 8.2) [119], entre la vérité terrain et les cartes de profondeur estimées à l'aide des trois méthodes.

$$PBM = \frac{1}{n} \times \sum_{x,y} D(x,y) \quad avec \quad D(x,y) = \begin{cases} 0 & si\ (|d_E(x,y) - d_T(x,y)| \leq \delta_d) \\ 1 & sinon \end{cases} \tag{8.2}$$

où :
- n est le nombre de pixels dans la carte de profondeur ;
- $d_E(x,y)$ et $d_T(x,y)$ sont respectivement la valeur de profondeur estimée du pixel (x,y) et celle de la vérité terrain ;
- δ_d est une valeur de tolérance. Dans nos expérimentations, nous utilisons $\delta_d = 1.0$ (valeur identique à celle utilisée dans [119]).

Une image de profondeur par personne dans la base Texas est utilisée, soit 118 images au total. Les mesures sont calculées sur toutes les images et la moyenne est rapportée. La Figure 8.5 montre que l'erreur RMS moyenne pour les 118 images est de 7,33 dans les résultats de la méthode BM et de 4,95 dans le résultat obtenu par notre méthode qui intègre le modèle de disparité dans le processus d'appariement. Concernant la valeur PBM, la Figure 8.6 montre que le pourcentage de faux appariements de pixels dans nos résultats est plus faible que celui obtenu en utilisant la méthode BM. De plus, par rapport à cette dernière, notre méthode nécessite moins de temps de calcul (voir la Figure 8.7, plus loin). Ceci est dû au fait que l'intervalle de disparité est fixé automatiquement à un petit segment de la ligne épipolaire en utilisant le modèle de disparité.

Les mesures RMS et PBM obtenues par notre méthode et la méthode GC sont proches. Cependant, comme le montre la Figure 8.7, notre méthode nécessite un temps de calcul 50

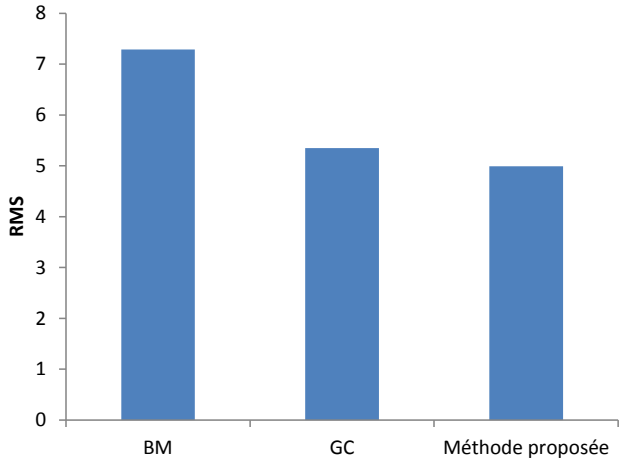

FIGURE 8.5: Mesure RMS moyenne pour les 118 personnes de la collection Texas.

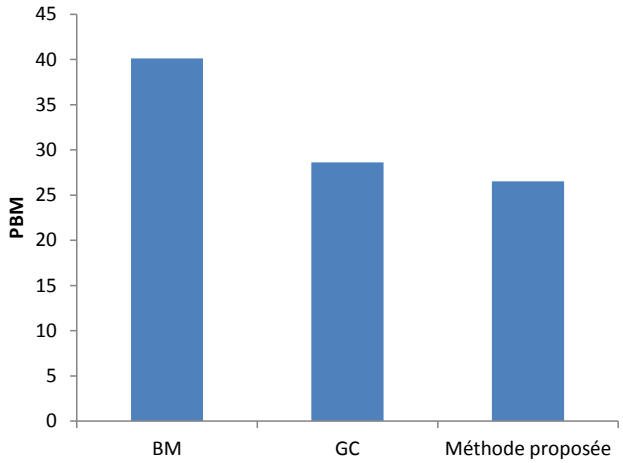

FIGURE 8.6: Mesure PBM moyenne pour les 118 personnes de la collection Texas.

fois plus petit que la méthode GC pour des images de 501×751 pixels. Nous pouvons donc constater que l'intégration des propriétés de forme du visage dans la procédure d'estimation améliore les résultats de l'estimation en termes de précision et de réduction de bruit dans les cartes de profondeur, tout en diminuant les temps de calcul.

Une dernière expérimentation effectuée sur cette collection a pour objectif de mesurer la ressemblance entre les visages estimés et ceux de la vérité terrain. Pour cela, nous construisons une matrice de similarité entre les modèles 3D reconstruits et les modèles 3D originaux (voir la Figure 8.8). Tout d'abord, nous avons projeté les pixels des cartes de profondeur dans un espace 3D afin de constituer un nuage de points 3D. Un maillage 3D simplifié est obtenu en reliant ces points. Ceci est effectué pour les cartes de profondeur de la vérité terrain et les cartes de profondeur estimées, obtenus à partir des trois méthodes. Ensuite, chaque maillage de la vérité terrain est comparé avec ceux obtenus avec notre méthode, la méthode BM et la méthode GC en utilisant l'algorithme ICP (*Iterative Closest Point*, voir la Section 3.3.1) [158]. Ceci permet de calculer l'erreur entre les maillages estimés et les maillages de la vé-

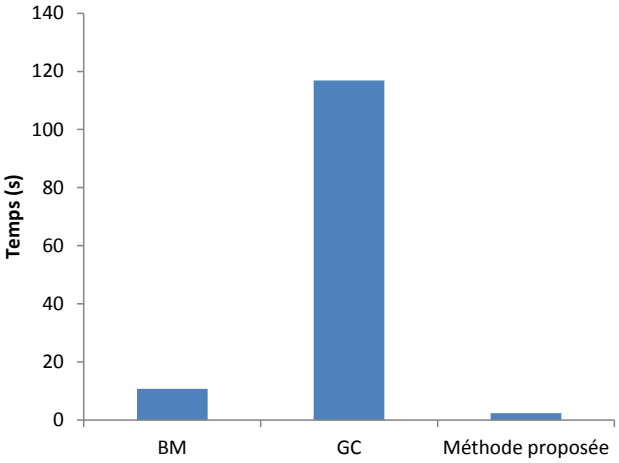

FIGURE 8.7: Temps de traitement.

rité terrain. La distance entre deux maillages est donnée par la moyenne des distances entre les points correspondants (*Mean of the Point-Wise Distance, MPWD*) [158] qui est calculée par l'Equation 8.3.

$$MPWD(m_T, m_E) = \frac{1}{n} \times \sum_{i=0, j=0}^{n-1} (D((P_i)^{m_T}, (P_j)^{m_E})) \tag{8.3}$$

où :

- m_T, m_E sont respectivement le maillage de la vérité terrain et le maillage reconstruit ;
- $D((P_i)^{m_T}, (P_j)^{m_E})$ est la distance calculée par ICP entre chaque paire de points correspondants ;
- n est le nombre de points utilisés dans l'algorithme ICP.

La Figure 8.8 montre les 3 matrices de similarités calculées entre la vérité terrain et les modèles 3D estimés, ainsi qu'une matrice calculée uniquement sur la vérité terrain (pour illustration). Chaque cellule (i, j) de la matrice représente la valeur $MPWD(m_t, m_e)$. Nous pouvons voir que la diagonale de la matrice obtenue par les résultats de notre méthode contient des valeurs inférieures (ligne claire) à celles obtenues par la méthode *block matching* et la méthode *graph cut*, ce qui signifie que les modèles reconstruits avec notre méthode sont plus précis et plus proches de leurs modèles 3D dans la vérité terrain. En outre, ces faibles valeurs ($\mu = 62,02mm$) sont différentes des autres valeurs plus élevées dans la matrice. Cette différence montre la spécificité des modèles 3D reconstruits par notre méthode, qui est assurée par l'utilisation d'un modèle de disparité pour chaque personne (obtenu en appliquant l'ASM) lors du calcul de la carte de profondeur.

8.3.3 Comparaison aux autres méthodes de reconstruction

. La comparaison de notre méthode stéréoscopique aux autres méthodes de reconstruction du visage est effectuée sur la collection Bosphorus [118]. Cette collection est plus riche

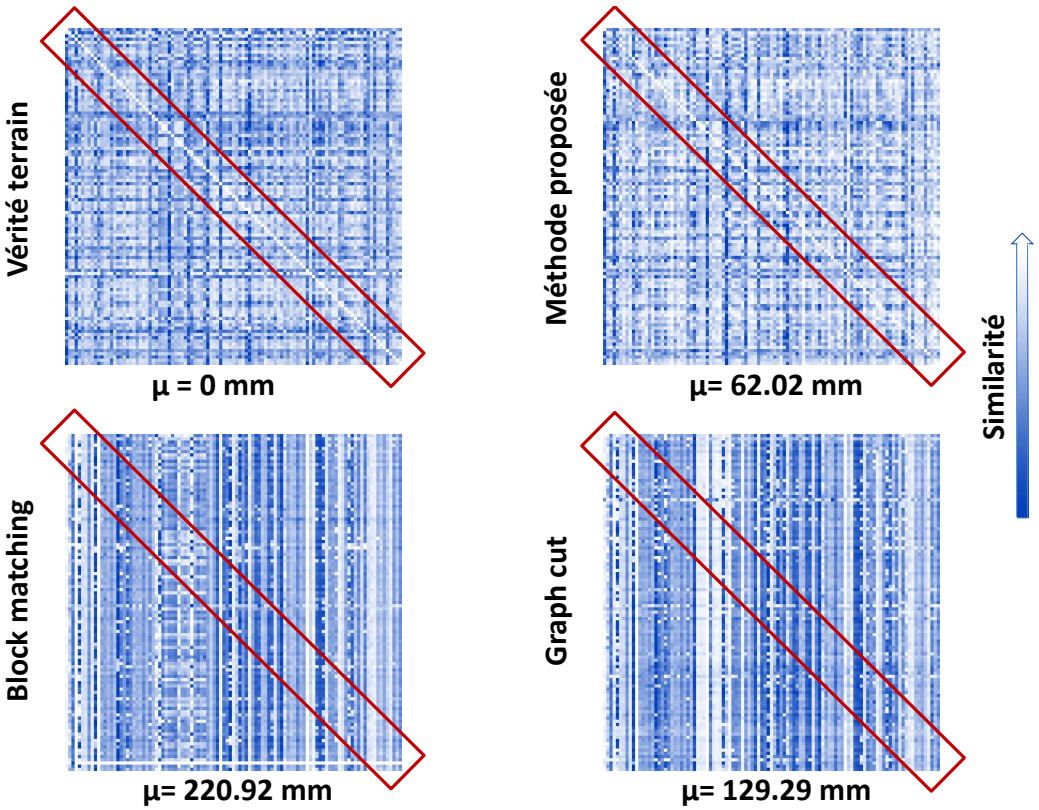

FIGURE 8.8: Matrices de similarité (μ est l'erreur moyenne de la diagonale).

en termes de variations d'expressions faciales et de pose. Cependant, les images ne présentent aucune variation d'éclairage. Elle contient 4652 images couleur de 105 personnes avec les images de profondeur correspondantes. En utilisant les images de profondeur de la vérité terrain de cette collection, les paires stéréoscopiques correspondant à ces images sont synthétisées afin d'appliquer notre méthode de reconstruction. L'évaluation sur collection Bosphorus est effectuée en utilisant cinq images avec des variations de pose différentes (annotées *PR_D*,*PR_SD*, *PR_SU*, *PR_U* et *YR_R*10) pour chaque personne. La Figure 8.9 montre des exemples des images utilisées. Afin de mesurer la précision des méthodes, un coefficient de corrélation est calculé entre le modèle reconstruit et le modèle de la vérité terrain en utilisant 22 points caractéristiques de la manière suivante :

$$\rho_{E,T} = \frac{cov(E,T)}{\sigma_E.\sigma_T} \tag{8.4}$$

où :
- E (resp. T) est l'ensemble des valeurs de profondeur estimées (resp. de la vérité terrain) pour les 22 points caractéristiques ;
- $cov(E,T)$ est la covariance de E et T ;
- σ_E (resp. σ_T) est l'écart type des valeurs de l'ensemble E (resp. T).

Dans la Figure 8.10, nous comparons nos résultats à ceux obtenus par Sun *et al.* dans

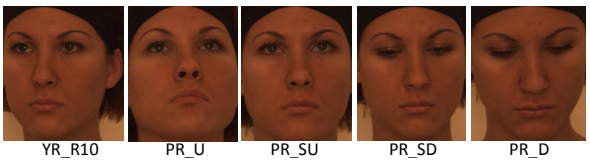

FIGURE 8.9: Exemples des images utilisées dans les expérimentations.

[127] sur un sous-ensemble des 30 premières personnes. Les auteurs proposent une méthode basée sur le modèle *CANDIDE-3* couplé avec la méthode *ICA* (*Independent Component Analysis*) ainsi qu'une extension de celle-ci où un processus d'intégration d'autres images est ajouté. Nous les nommons *ICA*0 et *ICA*1. Nous indiquons aussi les résultats de la méthode de Koo *et al.* [81] rapportés dans [127].

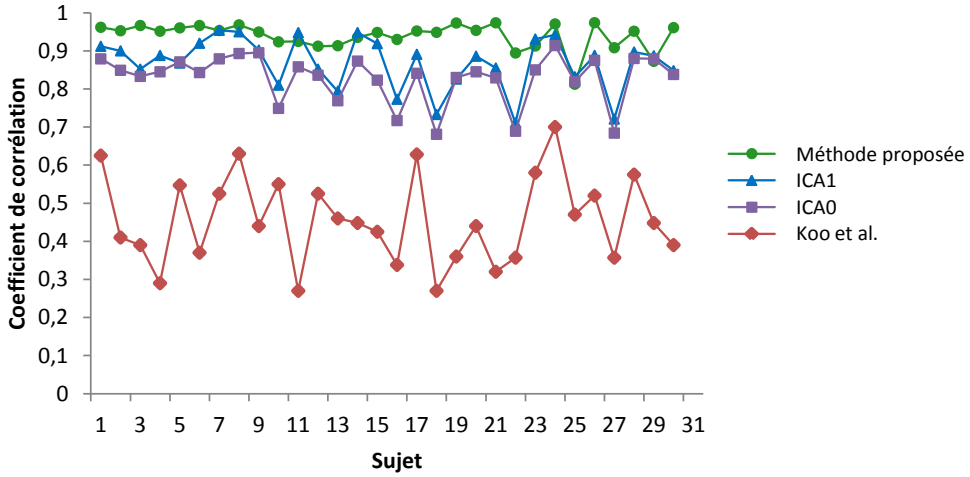

FIGURE 8.10: Coefficients de corrélation pour les 30 premières personnes de Bosphorus.

La Figure 8.10 montre les coefficients de corrélation des 30 première personnes de la collection Bosphorus. Nous pouvons voir que les résultats obtenus par Koo *et al.* [81] sont les moins précis, comparés à ceux obtenus par *ICA*0 et *ICA*1. Les résultats obtenus par *ICA*1 sont plus corrélés aux données de vérité terrain que ceux obtenus par *ICA*0. Cependant, notre méthode donne les meilleurs résultats de corrélation (toutes les valeurs sont proches de 1), ce qui prouve la précision du processus d'estimation proposé. En outre, les coefficients obtenus avec notre méthode sont stables, ce qui montre que notre méthode est plus robuste à l'identité de la personne que les autres méthodes, où les résultats varient fortement selon les personnes (en particulier la méthode de Koo *et al.*). Cela peut s'expliquer par le fait que nous utilisons un modèle de disparité spécifique pour chaque visage dans le processus d'estimation.

Dans la Figure 8.11, nous comparons nos résultats à ceux obtenus par Sun *et al.* [128]. Dans cette comparaison, seules 20 personnes ont été utilisées en raison de la disponibilité des résultats dans [128]. Les auteurs utilisent une technique de SfM (*Structure from Motion*) basée sur trois méthodes d'optimisation. La première est la méthode NLS (*Nonlinear*

Least Squares). Les deux autres sont des extensions de celle-ci, où les auteurs ont intégré la symétrie du visage et un processus de raffinement du modèle en utilisant d'autres images du visage. Nous nommons ici ces trois méthodes NLS0, NLS1 et NLS2. Les résultats de la méthode proposée par Fortuna *et al.* [50], rapportés dans [128], sont également présentés ici.

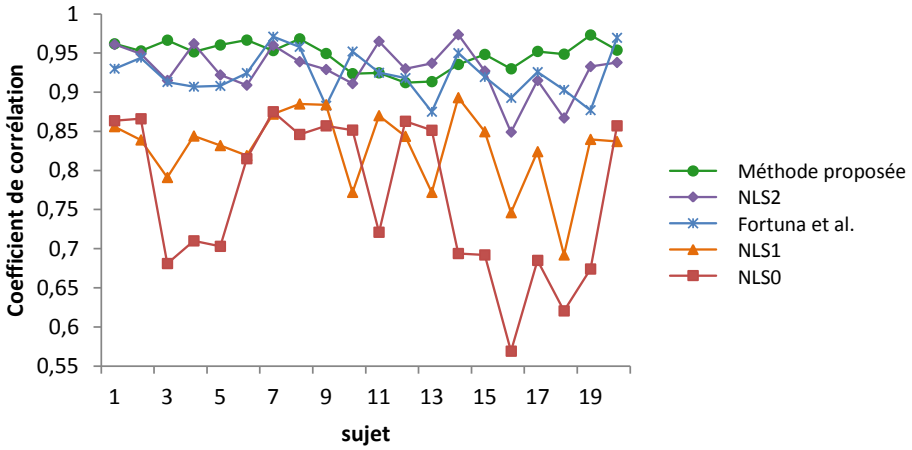

FIGURE 8.11: Coefficients de corrélation pour les 20 premières personnes de Bosphorus.

La méthode *NLS*0 donne les plus bas coefficients de corrélation et elle est très sensible aux changements de sujet. Les coefficients de corrélation sont légèrement améliorés en utilisant la propriété de symétrie du visage dans *NLS*1 et fortement améliorés lorsque plus de deux images sont utilisées par l'étape d'intégration de la méthode *NLS*3. Notre méthode donne des coefficients de corrélation très élevés pour tous les sujets et sont comparables à ceux obtenus par *NLS*2 et Fortuna *et al.* [50], qui sont basées sur un processus d'apprentissage où plusieurs images par sujet sont nécessaires. Notre méthode ne nécessite pas d'apprentissage, et ne nécessite qu'une seule paire d'images stéréoscopiques par personne.

Dans le Tableau 8.2, nous rapportons le nombre d'images nécessaires pour les méthodes de l'état de l'art utilisées pour la comparaison.

	Nombre d'images	Type des images
Koo *et al.* [81]	N/A	N/A
ICA0 [127]	2	Frontale + non frontale
ICA1 [127]	> 4	Frontale + non frontale
NLS2 [128]	> 4	Frontale + non frontale
Fortuna *et al.* [50]	N/A	N/A
Méthode proposée	2	paire stéréoscopique

TABLE 8.2: Nombre d'images utilisées par les méthodes de l'état de l'art et par notre méthode.

Dans le but de regarder en détails les valeurs de profondeur estimées, nous considérons la première personne comme exemple, et nous comparons les valeurs de profondeur estimées

pour 22 points caractéristiques aux valeurs de la vérité terrain (voir la Figure 8.12). Toutes les valeurs sont normalisées entre 0 et 1. La Figure 8.13 montre les écarts à la vérité terrain pour les 22 valeurs de profondeur estimées. Cette figure montre la précision de la profondeur estimée avec les différentes méthodes.

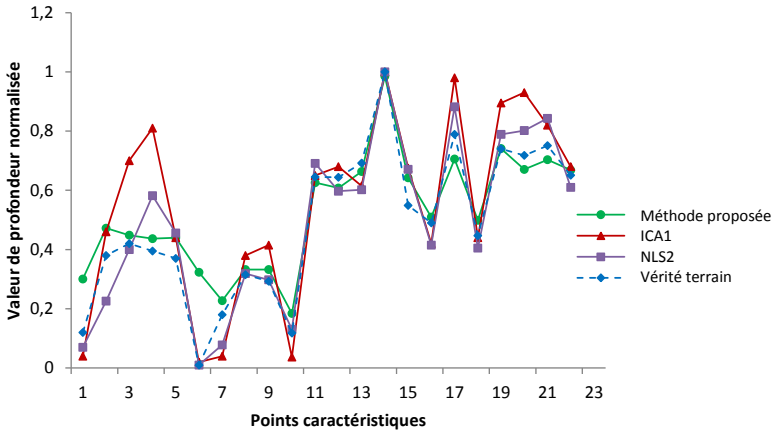

FIGURE 8.12: Valeurs de profondeur estimées et réelles pour les 22 points caractéristiques de la personne 1 de la collection Bosphorus.

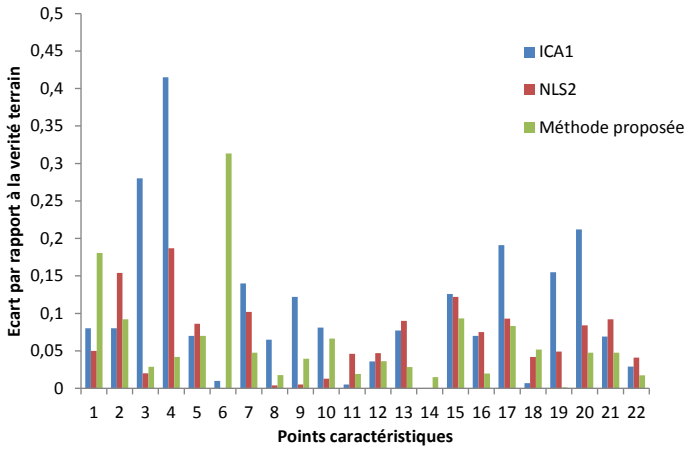

FIGURE 8.13: Écart entre les valeurs de profondeur estimées et réelles pour les 22 points caractéristiques de la personne 1 de la collection Bosphorus.

Afin d'évaluer les méthodes en cas de variations de pose, nous rapportons dans le Tableau 8.3 les coefficients de corrélation entre les valeurs de profondeur estimées et celles de la vérité terrain de cinq images de la personne 1 sous différentes poses (celles présentées à la Figure 8.9). Nous pouvons constater que les résultats de la méthode proposée sont élevés, et sont presque constants pour les différentes poses. La moyenne et l'écart type calculés montrent la précision et la robustesse de notre méthode.

	PR_D	PR_SD	PR_SU	PR_U	YR_R10	μ	σ
Koo *et al.* [81]	0.9312	0.2270	0.5665	0.7540	0.6201	0.6198	0.2608
ICA [127]	0.8822	0.8805	0.8775	0.8758	0.8789	0.8790	0.0025
NLS2 [128]	0.8916	0.8687	0.8380	0.8573	0.9015	0.8714	0.0257
Méthode proposée	**0.9678**	**0.9618**	**0.9478**	**0.9701**	**0.9616**	**0.9618**	**0.0057**

TABLE 8.3: Coefficients de corrélation pour des visages sous différentes poses.

Pour la commodité de l'affichage des résultats, seule une partie de la collection Bosphorus (20 et 30 personnes) est utilisée dans les expérimentations précédentes. Toutefois, afin d'évaluer les méthodes sur un grand nombre d'exemples, nous montrons dans le Tableau 8.4, la moyenne et l'écart type des coefficients de corrélation obtenus en utilisant les 105 personnes de la collection Bosphorus. Les valeurs des moyennes et des écarts types obtenus sur toutes les images de la collection confirment la précision de nos résultats, qui sont comparables aux meilleurs résultats obtenus par les méthodes de l'état de l'art.

	μ	σ
Koo *et al.* [81]	0.4920	0.2620
ICA1 [127]	0.8396	0.0631
ICA2 [127]	0.8708	0.0599
NLS2 [128]	0.9290	0.0313
Fortuna *et al.* [50]	0.9219	0.0290
Méthode proposée	0.9239	0.0261

TABLE 8.4: Moyenne et écart-type pour toutes les images de la collection Bosphorus.

8.4 Conclusion

Dans ce chapitre, nous avons évalué la méthode de reconstruction de visages proposée dans le cadre de cette thèse, indépendamment du processus de reconnaissance. D'abord la méthode proposée pour la débruitage des cartes de profondeur estimées est évaluée. La méthode proposée permet d'identifier les zones bruitées, elle ne nécessite pas de paramétrage et elle est invariante à la taille de la zone bruitée dans la limite où cette zone ne dépasse pas 50% de la totalité de la ligne de profondeur. En effet, si le bruit occupe plus de la moitié de la ligne de profondeur, l'identification est inversée. Nous notons par ailleurs que de telles situations sont peu probables dans les images reconstruites par notre méthode d'estimation de profondeur.

Différentes expérimentations ont été réalisées sur différentes collections afin d'évaluer la méthode d'estimation de profondeur. La comparaison de nos résultats à ceux obtenus par la méthode BM montre comment l'intégration du modèle de disparité permet l'amélioration du processus d'appariement en termes de précision et de temps de calcul. Nos résultats

sont comparables aux résultats obtenus avec la méthode GC, tout en nécessitant un temps de calcul moindre. Les résultats obtenus montrent l'efficacité de notre méthode pour la reconstruction stéréoscopique de visages.

Le dernier chapitre de cette partie est consacré à l'évaluation des deux autres contributions proposées dans le cadre de cette thèse : la description des images de profondeur pour la représentation des visages et la fusion des deux modalités 2D et 3D.

Chapitre 9

Évaluation de la méthode bimodale de reconnaissance de visages

Sommaire

9.1 Introduction 108
9.2 Collections de tests 108
9.3 Évaluation du descripteur DLBP 109
9.3.1 Étude des paramètres des DLBP 109
9.3.2 Comparaison avec les autres descripteurs 111
9.4 Évaluation de l'approche globale de reconnaissance bimodale 114
9.5 Conclusion 116

9.1 Introduction

Dans ce chapitre, nous évaluons qualitativement et quantitativement les résultats de l'approche bimodale de reconnaissance de visages proposée, indépendamment de la technique utilisée pour l'acquisition des données 3D. Tout d'abord, une évaluation du descripteur proposé pour les images de profondeur, le DLBP, est effectuée sur différentes collections de tests de visages 3D. Ensuite, différentes expérimentations sont réalisées en vue d'évaluer les deux modalités 2D et 3D séparément et comparer leurs performances à celles obtenues avec la méthode bimodale de reconnaissance. Les trois stratégies de fusion (la fusion de descripteurs, de décisions et bi-niveaux) sont mise en œuvre et comparées. Des collections de différentes résolutions obtenues par plusieurs capteurs (caméra stéréoscopique, scanner 3D et caméra à infrarouge de type Kinect) sont utilisées.

9.2 Collections de tests

Les expérimentations présentées dans ce chapitre sont effectuées sur six collections :

- **FRGC** [110] : la collection de visages 3D FRGC est la plus utilisée dans l'état de l'art. Elle est composée de 4007 images couleur de 446 personnes avec les images de profondeur correspondantes. Dans cette collection, les images sont acquises en faisant varier l'illumination (différentes conditions d'éclairage) et les expressions faciales. La totalité des images de visages est prise de face. Nous soulignions que nous avons retiré un sous-ensemble d'images (environ 15%) du fait que les cartes de profondeur correspondantes ne sont pas exploitables à cause de la présence de grands trous dans les modèles 3D.
- **Texas** [56] : cette collection contient un total de 1149 images couleur de 118 personnes, ainsi que les images de profondeur correspondantes. Toutes les prises de vue sont de face. Pour chaque personne, la variabilité des images provient principalement de la variation d'éclairage, et dans une moindre mesure, de quelques variations des expressions faciales. Cette collection est la même que celle utilisée dans la Section 8.3.2.
- **Bosphorus** [118] : cette collection est plus riche en termes de variations d'expressions faciales et de pose. Cependant, les images ne présentent aucune variation d'éclairage. Elle contient 4652 images couleur de 105 personnes avec les images de profondeur correspondantes. Cette collection a également été utilisée dans la Section 8.3.3.
- **TexasStereo** : Cette collection contient les mêmes données que Texas. La seule différence est que les cartes de profondeur sont obtenues en utilisant notre méthode de reconstruction stéréoscopique proposée (cf. Section 8.3.2).
- **FoxStereo et FoxKinect** : ce sont deux collections construites dans le cadre de cette thèse, décrites dans le Chapitre 7 [1]. Les cartes de profondeur de la collection FoxStereo sont obtenues en utilisant la méthode de reconstruction proposée, à partir des images

1. La collection FoxTOF n'a pas été utilisée dans nos expérimentations du fait qu'elle contient des images proches de l'infrarouge (pas couleur) non exploitable dans le cadre de notre approches bimodale.

droites et gauches des visages.

Prétraitement

Un processus de prétraitement a été appliqué sur toutes les collections. Les visages sont d'abord extraits à partir des images en niveaux de gris et de profondeur, en utilisant les fichiers d'annotation fournis. La taille des images de visages est ensuite normalisée à 100×100 pixels. Enfin, le débruitage des cartes de profondeur (décrit dans la Section 5.3) de toutes les six collections est effectué. Dans nos expérimentations, nous n'avons considéré que les images de visages présentant une pose frontale. Ainsi aucun alignement de pose n'a été nécessaire.

9.3 Évaluation du descripteur DLBP

Dans cette section, nous décrivons l'évaluation du descripteur proposé sur différentes collections de visages 3D. Un prototype de reconnaissance faciale est mis en œuvre afin de comparer la précision des DLBP à celle obtenue par le descripteur LBP classique [5] et le descripteur 3DLBP [69, 65]. Les valeurs brutes des pixels sont aussi utilisées dans la comparaison afin de montrer le gain en précision obtenu par rapport à l'utilisation directe des cartes de profondeur. La méthode du plus proche voisin (1NN) a été utilisée pour la classification. Cette méthode est simple et permet de comparer le pouvoir discriminant des descripteurs indépendamment du processus d'apprentissage. Les taux de reconnaissance sont obtenus par la méthode de validation croisée en 10 plis (*10-fold cross validation*). Le prototype utilisé permet une comparaison des différents descripteurs indépendamment du prétraitement utilisé et de la méthode de classification adoptée. Le processus de reconnaissance a été effectué sur les six collections décrites dans la section précédente.

9.3.1 Étude des paramètres des DLBP

Nous étudions dans cette section le comportement des DLBP, par rapport aux changements de paramètres, notamment : le rayon R et le nombre d'histogrammes élémentaires extraits H (qui dépends de la taille de la grille utilisée pour découper l'image en régions). La taille du vecteur DLBP augmente d'une façon exponentielle par rapport à la taille du voisinage V (i.e. le nombre de voisins considérés). En effet, la taille du descripteur est de $2^V \times H$. Le regroupement des codes LBP dans un même *bin* permet la réduction de la taille du descripteur. Cependant, les codes de LBP sont indépendants, et alors il ne serait pas raisonnable d'envisager de les regrouper dans des *bins* plus gros afin de réduire la dimension comme cela est fait classiquement pour les histogrammes de couleur. Par conséquent, nous fixons la taille du voisinage V à 8 dans toutes les expérimentations.

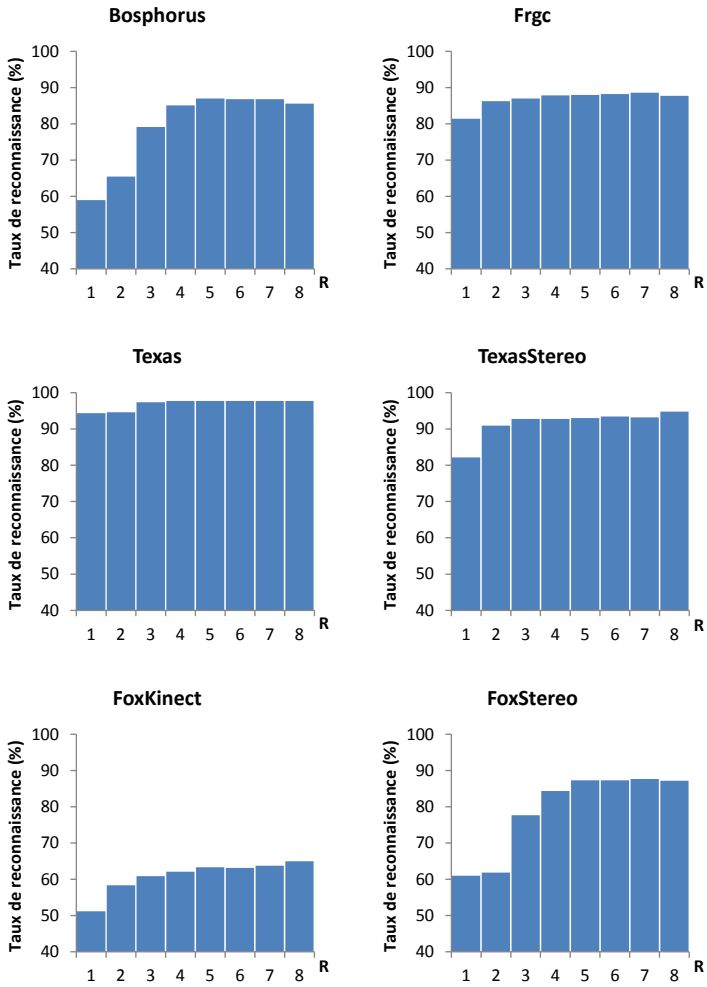

FIGURE 9.1: Impact de la variation du paramètre R sur la précision des DLBP ($H = 25$).

Impact de R

La Figure 9.1 présente une comparaison des taux de reconnaissance obtenus avec différentes valeurs de rayon R. Elle montre que les résultats sont globalement bons sur les différentes collections de tests utilisées. Le changement du rayon R contribue significativement à l'amélioration de la performance du système de reconnaissance. En effet, plus le rayon est grand (jusqu'à certain seuil), plus le taux de reconnaissance est élevé. Ceci valide notre intuition qui préconise de considérer un grand voisinage lors de l'application des DLBP aux images de profondeur de visages (cf. Section 6.3.2). En effet, le contraste en profondeur dans un voisinage petit n'est pas suffisant pour la discrimination à cause de l'aspect lisse de la surface du visage. Par ailleurs, nous pouvons voir que l'impact de R change pour les différentes collections. Ceci peut être expliqué pour deux raisons. La première raison est le fait que les collections contiennent des cartes de profondeur de différentes résolutions. Ainsi la variation du rayon ne se traduit pas de la même façon pour toutes les collections. La deuxième rai-

son qu'on peut considérer est la marge d'amélioration possible. En effet, lorsque le taux est grand, l'amélioration qui peut être apportée est petite.

Impact de H

La comparaison des performances des DLBP par rapport au changement du nombre d'histogrammes est illustré dans la Figure 9.2. Nous pouvons remarquer que :

- les résultats obtenus pour toutes les collections de tests sont globalement bons (entre $97,73\%$ pour Texas et $70,49\%$ pour FoxKinect de taux de reconnaissance).
- les résultats confirment que le taux de reconnaissance augmente avec le rayon et se stabilise à une valeur entre 4 et 5 pour la plupart des collections.
- le nombre d'histogrammes utilisés influence la performance des DLBP. Plus le nombre d'histogrammes H par image est grand, plus le taux de reconnaissance est élevé. $H = 25$ donne des résultats souvent meilleurs. Par ailleurs, plus H augmente, plus le gain se réduit. On peut supposer un gain asymptotique, et une plus grande valeur de H ajouterait plus de complexité (dimensionnalité, temps d'exécution de 1NN) pour un gain réduit.
- pour toutes les collections, plus le rayon est grand, plus l'impact de H est petit. Ce comportement trouve une explication dans le faible contraste de profondeur des visages. En effet, quand le rayon est petit, des zones différentes du visage sont susceptibles d'être représentées par des codes identiques. Dans le cas d'un histogramme unique ($H = 1$), il en résulte un grand nombre de codes distribués dans peu de *bins*. Par ailleurs, utiliser plus d'histogrammes ($H > 1$) aide à préserver localement l'information spatiale de l'emplacement des codes. Quand R est grand, à l'inverse, le contraste de profondeur devient plus aisément distinguable par les codes (les codes sur une région donnée se ressemblent moins), et le nombre de codes différents est plus grand, ce qui produit un descripteur plus discriminent, même pour $H = 1$. C'est pour cela que l'écart de précision entre les différentes valeurs de H est réduit.

L'étude expérimentale effectuée pour l'évaluation des DLBP démontre les deux points suivants :

- le paramètre R est important pour le pouvoir discriminant des DLBP. En effet, l'utilisation des DLBP à grande échelle a permis d'obtenir les meilleurs taux de reconnaissance ;
- la précision des DLBP par rapport au nombre d'histogrammes utilisés lors de l'utilisation d'une grande valeur de rayon est relativement stable pour la majorité des collections de test utilisées.

9.3.2 Comparaison avec les autres descripteurs

Après avoir évalué les DLBP, dans la section précédente, nous présentons dans cette section une étude comparative entre notre descripteur et d'autres descripteurs proposés pour la représentation des images de profondeur. Nous comparons dans la Figure 9.3 les taux de reconnaissance obtenus en utilisant des histogrammes des valeurs de profondeur brutes à

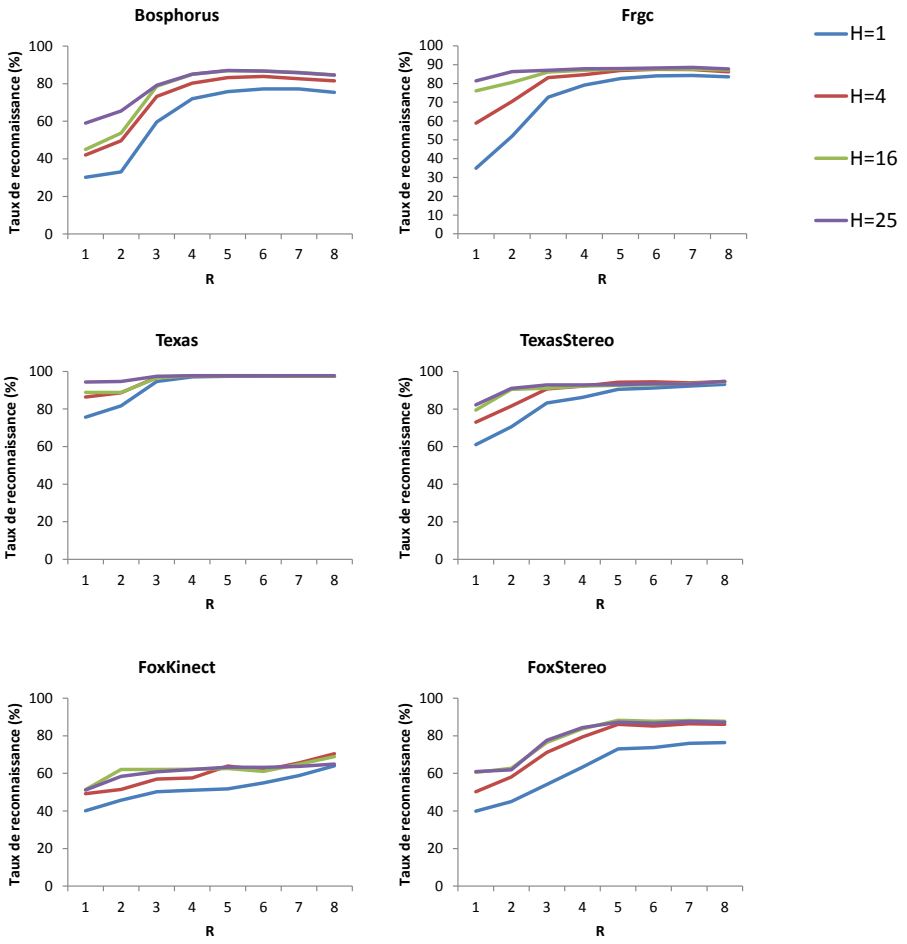

FIGURE 9.2: Impact de la variation du paramètre H sur la précision des DLBP.

ceux obtenus en appliquant les DLBP. Un ensemble de 25 histogrammes de 256 *bins* sont extraits pour les deux ensembles de données (pixels et DLBP). Différentes valeurs de rayon ont été utilisées, et la performance maximale sur chaque collection a été rapportée.

Nous pouvons remarquer que l'utilisation des DLBP permet d'améliorer la performance du système de reconnaissance par rapport à l'utilisation des valeurs de profondeur brutes (les pixels de la carte de profondeur). En effet, les DLBP peuvent faire ressortir des contrastes en profondeur qui permettent une meilleure discrimination entre les cartes de profondeur des différents visages.

Dans la Figure 9.4, nous comparons les DLBP aux LBP et aux 3DLBP, sur les six collections de visages. Nous effectuons nos expérimentations selon la version originale du 3DLBP, proposée par Yonggang Huang *et al.* [69], appliquée avec un rayon $R \in \{1,2\}$, ainsi qu'à différentes échelles comme il a été proposé par Di Huang *et al.* [65]. Le nombre d'histogrammes utilisé est 25. Nous pouvons voir que les DLBP et les 3DLBP donnent de bons résultats par rapport à l'utilisation du LBP classique. Ceci montre que l'utilisation de la magnitude contribue significativement dans l'amélioration des taux de reconnaissance. Nous notons aussi une

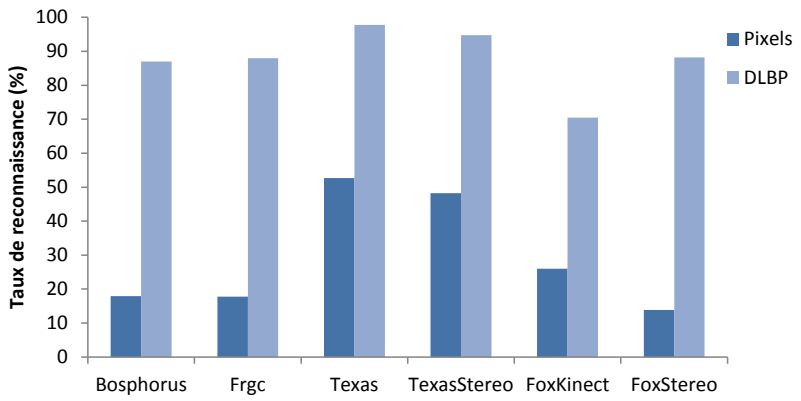

FIGURE 9.3: Comparaison entre l'utilisation des valeurs de profondeur brutes et les DLBP.

amélioration des taux de reconnaissance en augmentant le rayon. Cette amélioration est plus significative pour les LBP et les DLBP. En effet, l'application des 3DLBP sur grande échelle ne permet pas d'améliorer la précision de reconnaissance car le schéma de codage sur lequel ils sont basés ne permet pas d'exploiter le contraste de profondeur à grande échelle.

La comparaison des DLBP à l'utilisation des histogrammes des valeurs de profondeur brutes, en premier lieu et aux LBP classiques, aux 3DLBP [69] et à leur extension multi-échelles [65], en deuxième lieu, montre quatre points essentiels :

- l'utilisation des DLBP permet une meilleure discrimination que l'utilisation directe des valeurs des pixels ;
- la précision des descripteurs DLBP et 3DLBP est meilleure que celle obtenue à l'aide du LBP classique. Ceci montre l'importance de l'utilisation de l'information de magnitude en plus du signe dans la description de la forme 3D du visage ;
- le descripteur DLBP permet d'atteindre une meilleure précision que les deux autres descripteurs. Ceci est dû à l'exploitation cohérente de l'information de magnitude, ainsi qu'à l'application du seuillage automatique de cette dernière en considérant les gradients multi-échelles, ce qui a permis d'étendre le descripteur aux grandes échelles. En effet, à grande échelle, les différences du voisinage semblent être plus discriminantes, étant donné l'aspect lisse de la forme 3D du visage. L'hypothèse de base supposant que les différences de voisinage ne dépassent pas la valeur 7, sur laquelle le descripteur 3DLBP est basé, limite son utilisation à un rayon inférieur à 2 et par conséquent son pouvoir discriminant est inférieur. Les auteurs de [65] ont proposé de calculer le 3DLBP à différentes échelles, comme nous l'avons effectué dans les expérimentations présentées ci-dessus. Cependant, ceci n'a pas permis d'améliorer la précision. En effet, les auteurs ont gardé le même principe du 3DLBP et donc les magnitudes à grande échelle sont seuillées par la valeur 7 fixée par le 3DLBP ;
- en plus d'un pouvoir de discrimination meilleur, la taille du descripteur DLBP est 2 fois plus petite que le descripteur 3DLBP.

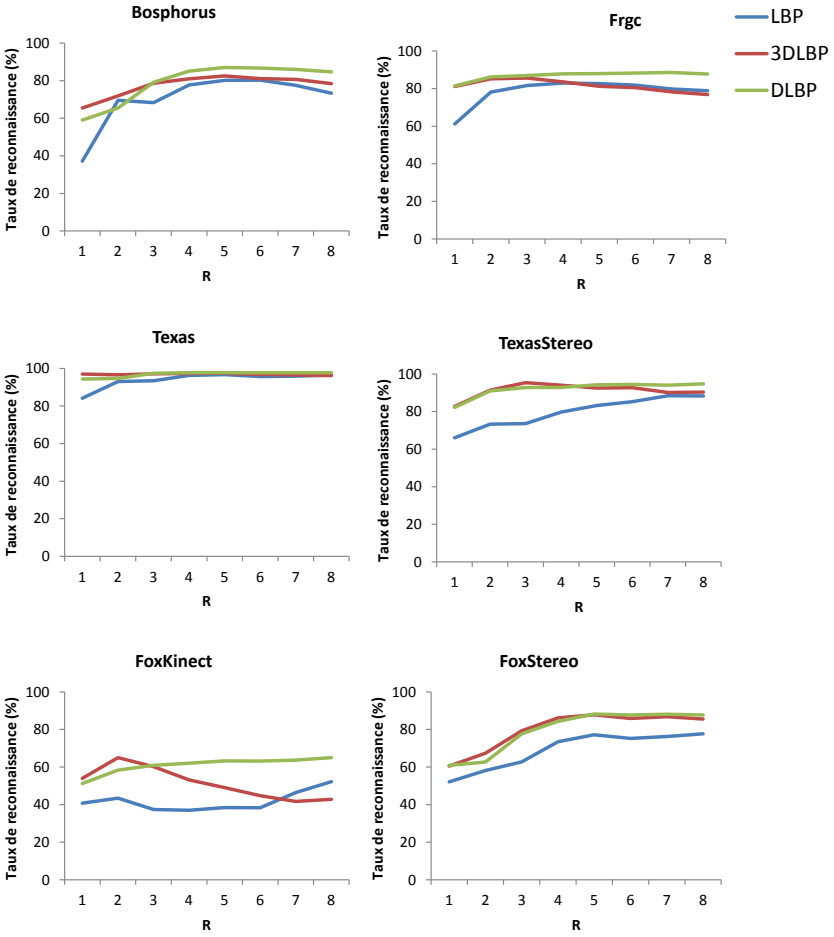

FIGURE 9.4: Comparaison entre DLBP, LBP et 3DLBP.

9.4 Évaluation de l'approche globale de reconnaissance bimodale

Dans cette section, nous évaluons la qualité globale de l'approche de reconnaissance bimodale proposée dans le cadre de cette thèse pour la reconnaissance bimodale. Après avoir évalué précédemment le descripteur proposé pour la représentation des images de profondeur, nous évaluons dans cette section l'approche globale incluant l'étape de classification, ainsi que les différentes stratégies de fusion adoptées. L'apprentissage repose sur des machines à vecteur de support (*Support Vector Machines, SVM*), avec une fonction à noyau radial (*Radial Basis Function, RBF*). Nous avons choisi les SVM pour leur efficacité démontrée plusieurs fois dans l'état de l'art. Pour évaluer la précision, une validation croisée en 10 plis (*10-fold cross validation*) est appliquée. Les paramètres qui ont donné les meilleurs taux de reconnaissance dans les expérimentations précédentes sont utilisés pour cette étude comparative. Les taux de reconnaissance obtenus sur les 6 collections utilisées précédemment sont indiqués.

Dans la Figure 9.5, nous comparons les deux approches monomodales (2D et 3D) aux approches bimodales basées sur les trois stratégies de fusion (descripteurs, décisions et bi-niveaux). Nous rappelons que pour la fusion des descripteurs, les deux vecteurs caractéristiques des deux modalités 2D et 3D sont concaténés afin d'en construire un seul. Par ailleurs, la fusion de décision et la fusion bi-niveaux sont effectuées par la méthode de vote majoritaire pondéré. Les résultats montrent cinq points importants :

- dans les collections FRGC et Texas, la reconnaissance 3D donne une meilleure précision par rapport à la reconnaissance 2D, contrairement aux autres collections. Ceci s'explique par le fait que les variations d'éclairage dans ces deux collections sont très importantes. Par conséquent, l'utilisation de la profondeur permet une discrimination plus robuste, et donc de meilleurs taux de reconnaissance sont obtenus. Bien que la collection TexasStereo contienne les mêmes variations, les images de profondeur étant d'une qualité plus basse (images reconstruites), elles ne permettent pas d'obtenir de meilleurs résultats que l'approche 2D ;
- la fusion de descripteurs améliore légèrement la performance du système pour les collections FRGC et Texas. Cependant, ceci ne s'applique pas sur le reste des quatre collections. Dans la collection Bosphorus par exemple, la fusion de descripteurs donne une sorte de moyenne des taux obtenus avec les deux modalités 2D et 3D utilisées séparément. Ceci peut être expliqué par la relation de complémentarité. En effet, la fusion de descripteurs est plus appropriée quand les données sont complémentaires, ce qui est le cas dans les collections Texas et FRGC où les variations d'éclairage sont importantes. Une autre explication possible est la qualité des données de la collection. En fusionnant les données, nous accumulons aussi le bruit éventuellement présent dans les vecteurs des deux modalités. Par conséquent, si un vecteur est très bruité, il influence la totalité du descripteur fusionné, et diminue ainsi la performance globale du système ;
- la fusion de décisions améliore ou conserve les performances du système pour toutes les collections. Ceci montre que la fusion de décisions est une meilleure option que la fusion de descripteurs, du fait que chaque descripteur est traité séparément. Si un des deux descripteurs est meilleur que l'autre pour une classe donnée, sa décision ne sera pas influencée par l'autre descripteur ;
- la fusion bi-niveaux s'avère être meilleure que les deux autres stratégies utilisées séparément. Elle donne une précision supérieure à celles obtenues par les fusions de descripteurs ou de décisions dans la plupart des collections et elle garantit la même précision que la fusion de décisions lorsqu'aucune amélioration n'est apportée ;
- un point important qui doit être souligné est l'impact de la résolution des cartes de profondeur sur à la précision de la reconnaissance. Nous pouvons remarquer que les données de profondeur dans TexasStereo sont de moins bonne résolution que celles de la collection Texas. Dans TexasStereo, les données sont obtenues à l'aide de notre méthode de reconstruction et ne sont pas de la même qualité que les cartes de profondeur originales de Texas. Cependant, la fusion de ces données avec les données 2D permet d'augmenter la précision du système.

Nous pouvons constater que d'une façon générale, l'approche bimodale proposée permet

d'améliorer la précision de la reconnaissance de visage. En se basant sur les expérimentations effectuées sur les différentes collections, nous considérons que la fusion de décisions et bi-niveaux sont plus précises par rapport à la fusion de descripteurs. La stratégie de fusion bi-niveaux permet une amélioration de la précision de la reconnaissance. Elle garantit aussi la précision maximale obtenue par la fusion de décisions si aucune amélioration n'est possible. En effet, aucune dégradation de précision n'a été obtenue même dans le cas où la fusion de descripteurs est moins précise. Ceci est dû au fait que les décisions utilisées par la fusion bi-niveaux (obtenues par l'approche 2D, l'approche 3D et l'approche de fusion de descripteurs) sont prises indépendamment les unes des autres. Si un des deux descripteurs est meilleur que l'autre pour une classe donnée, la décision de chaque classifieur correspondant n'est pas influencée par l'autre.

9.5 Conclusion

Dans ce chapitre, nous avons présenté une validation expérimentale de deux des contributions de ce travail de thèse. Six collections de visages de différentes tailles contenant différentes variations sont utilisées. Dans un premier temps, le descripteur proposé pour la description des cartes de profondeur a été évalué sur différents paramètres et une comparaison avec les méthodes de l'état de l'art a été effectuée. Les résultats montrent trois points essentiels. Le premier point est l'intérêt de l'utilisation de l'information de magnitude en plus du signe pour la description de la profondeur. Le deuxième consiste en la validation de notre intuition qui indique que l'utilisation d'un voisinage à grande échelle permet de faire ressortir le contraste de profondeur permettant ainsi une meilleure discrimination des visages. Le dernier point est que notre descripteur donne les meilleures performances tout en ayant une taille plus compacte que celle des 3DLBP. Une étude comparative entre les approches monomodales et bimodales d'une part, et entre différentes stratégies de fusion d'une autre part, est ensuite effectuée. La première conclusion que l'on peut tirer des résultats est que la fusion des deux modalités permet une amélioration des performances des approches monomodales. De plus, la stratégie bi-niveaux permet de tirer profit des deux stratégies de fusion (tardive et précoce) et permet de donner un taux maximum de reconnaissance. Ce chapitre conclut la partie expérimentale de notre travail. La partie suivante donne une conclusion générale du travail effectué dans le cadre de cette thèse, ainsi que les perspectives envisagées.

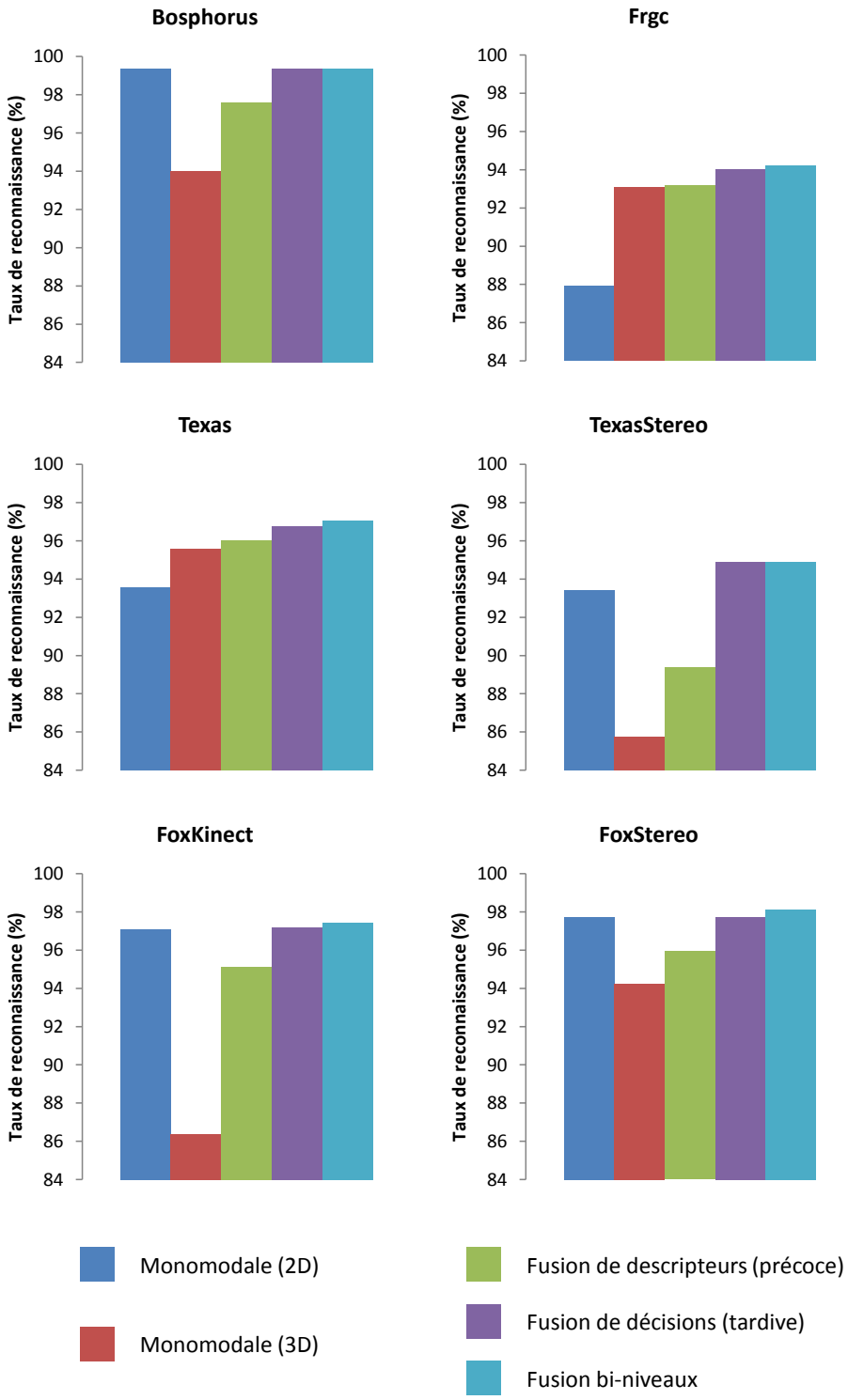

FIGURE 9.5: Comparaison entre les méthodes monomodales (2D et 3D) et les méthodes bimodales basées sur différentes stratégies de fusion.

Quatrième partie

Conclusion générale

Chapitre 10

Conclusion

Sommaire

10.1 Synthèse des contributions 122
10.2 Perspectives 123

10.1 Synthèse des contributions

Dans cette thèse nous nous sommes intéressés aux approches bimodales 2D-3D de reconnaissance de visages. La combinaison des indices visuels 2D et de profondeur conduit à améliorer la reconnaissance faciale, car une telle approche exploite les avantages et la complémentarité des deux modalités et peut ainsi permettre de combler leurs lacunes respectives.

Tout au long de cette thèse, des solutions aux différents problèmes rencontrés dans le processus de reconnaissance bimodale de visages sont proposées. Nous avons traité trois problématiques majeures : l'acquisition 3D du visage, la représentation efficace des données 3D et la fusion des deux modalités 2D et 3D.

- **Acquisition des données 3D** : concernant ce point, notre contribution consiste en la proposition d'une méthode originale d'**estimation de la profondeur** du visage dans un système stéréoscopique. À la différence des méthodes générales utilisées pour le calcul de la disparité des objets quelconques, nous avons introduit une méthode spécifique pour l'estimation de la profondeur du visage qui utilise des caractéristiques topologiques liées à la forme du visage humain, obtenues par une étape d'ajustement d'un ASM, afin d'améliorer les résultats de l'estimation. Les résultats expérimentaux ont montré que notre proposition améliore en termes de précision la méthode de *block matching* classique pour le calcul des disparités, tout en permettant un traitement plus rapide. L'algorithme proposé produit des cartes de profondeur denses et lisses de visages, utilisables dans de nombreuses applications.
- **Représentation des données 3D :** ce point constitue la deuxième contribution de notre travail. Nous avons proposé un **descripteur nommé *DLBP*** pour la représentation des cartes de profondeur. Ce descripteur est inspiré des LBP qui ont connu un grand succès pour la reconnaissance 2D de visages. Le descripteur proposé prend en compte la nature des données 3D par rapport aux données visuelles 2D, en intégrant les valeurs de la magnitude des différences de profondeur et en considérant le voisinage à grande échelle. De plus, il est basé sur une étape de seuillage automatique qui tient compte de la résolution des cartes de profondeur utilisées. Il permet donc une meilleure représentation de la profondeur du visage, ainsi qu'une certaine flexibilité concernant la résolution des images de profondeur (en comparaison avec les méthodes géométriques utilisées dans l'état de l'art nécessitant des données 3D de haute résolution). Les expérimentations effectuées afin d'évaluer les DLBP ont montré leur robustesse et leur pouvoir discriminant sur différentes collections de tests.
- **Fusion 2D-3D :** la troisième contribution de ce travail de thèse consiste en la **fusion des données 2D-3D** afin d'améliorer la précision et la robustesse de la reconnaissance par rapport à l'utilisation de chaque modalité seule. Nous avons envisagé deux stratégies de fusion : précoce et tardive. Nous avons ainsi proposé une stratégie de fusion bi-niveaux afin de bénéficier pleinement des avantages des deux niveaux de fusion. Les résultats expérimentaux ont montré que la stratégie proposée permet d'augmenter la précision de la reconnaissance et de conserver la performance maximale obtenue par les deux niveaux (précoce et tardif) dans le cas où aucune amélioration n'est apportée. Les expérimentations effectuées ont aussi montré que l'utilisation conjointe des

modalités 2D et 3D permet d'améliorer notablement la précision de reconnaissance même si la résolution des données 3D utilisées est basse. Il est important de noter que cette amélioration est constatée bien que la précision atteinte à l'aide de la modalité 3D seule baisse avec la résolution et la qualité. Cela démontre que des données de profondeur de basse résolution peuvent améliorer la précision de reconnaissance dans un contexte de reconnaissance bimodale.

Par ailleurs, dans le cadre de ce travail de thèse, nous avons été amenés à élaborer une collection de tests pour répondre aux besoins de l'évaluation de nos propositions. Cette collection comprend les enregistrements visuels (images 2D et vidéos) et de profondeur (acquis à l'aide de 3 dispositifs différents notamment une caméra stéréoscopique, une caméra à infrarouge et une caméra ToF) de 64 personnes présentant des variations d'éclairage, de pose, et d'expressions faciales. Nous soulignons que cette collection peut aussi être utilisée pour l'évaluation de différentes méthodes d'analyse de visages (reconnaissance de visages, reconnaissance d'émotions et de la pose, reconstruction de visages).

Les travaux réalisés dans ce travail de thèse ont permis d'atteindre l'objectif que nous nous sommes fixé et qui était l'origine de ce travail de thèse.

10.2 Perspectives

Ce travail donne lieu à de nombreuses perspectives :

- **Au niveau reconstruction** : notre approche d'estimation de profondeur du visage ouvre de nombreuses perspectives d'amélioration et d'extension. L'étape de la reconstruction du modèle de disparité peut être incorporée dans d'autres méthodes d'appariement stéréoscopiques afin de réduire le temps de traitement et le nombre de fausses correspondances. Dans le cas de la méthode *graph cut* par exemple, l'étape d'optimisation sur laquelle se base cette méthode cherche à trouver une coupure optimale d'un graphe représentant toutes les disparités possibles pour chaque pixel de l'image. L'intégration de l'étape de construction du modèle de disparité proposée permet de réduire l'intervalle de disparité pour chaque pixel ce qui permettrait de réduire considérablement le temps de traitement de la méthode. L'estimation du modèle de disparité peut également être améliorée en utilisant les modèles actifs d'apparence 3D [154], qui sont plus robustes que les ASM aux variations de poses.
- **Au niveau description de données 3D** : il serait intéressant d'étendre notre approche de reconnaissance pour qu'elle soit invariante aux changements de pose. En effet, l'utilisation des cartes de profondeur limite l'invariance des données 3D à la pose par rapport à l'utilisation d'un maillage 3D complet du visage. Ceci est dû à la projection des données 3D sur un plan 2D. Nous projetons donc d'explorer deux voies afin de faire face à ce problème. La première consiste à modifier le mode de fonctionnement des DLBP afin qu'il soit invariant aux petites variations de la surface du visage selon les axes X (*pitch* ou *tilt*, ou tangage) et Y (*yaw* ou *pan*, ou lacet). Cela demande de traduire ces rotations de surface au niveau des pixels de la carte de profondeur afin de modifier l'étape de construction des DLBP. La deuxième voie est d'intégrer un processus de

prétraitement consistant à estimer la pose afin de la corriger avant l'application des DLBP. La correction des petites variations de pose ($\leq 20°$) peut être effectuée au niveau local, cependant pour des grandes variations ($> 20°$), on peut envisager de passer par une reconstruction 3D, dans le but de la projeter sous une pose frontale.

– **Au niveau fusion** : la fusion des données 2D-3D peut être améliorée en s'appuyant sur un cadre probabiliste ou bayésien [30]. En effet, de nombreuses méthodes de fusion s'appuient sur ce cadre dédié. Ceci doit permettre de tenir compte des imprécisions et des incertitudes lors du processus de la fusion. Un autre point à explorer dans cette direction est la fusion des deux modalités, non seulement aux deux niveaux sur lesquels nous nous sommes concentrés dans le cadre de notre travail (descripteurs et de décisions), mais aussi au niveau du prétraitement des données. En effet, les données 2D et 3D semblent être complémentaires pour différents processus de prétraitement comme par exemple la détection des points caractéristiques. L'avantage est que la position des points caractéristiques du visage est la même pour les deux types de données. Les données 2D peuvent être donc utilisées pour localiser les yeux alors que les données 3D sont plus utiles pour la détection du bout du nez. L'estimation et la correction de la pose est un autre exemple de prétraitement où la fusion peut être exploitée. En effet, les données 3D sont plus appropriées pour estimer la pose (en utilisant le bout du nez) et les images 2D sont utiles pour texturer le modèle une fois la pose corrigée afin d'obtenir une nouvelle image 2D sous une pose frontale. Une autre piste intéressante que nous envisageons afin d'améliorer la fusion est l'étude de la dépendance des données 2D-3D. Pour cela, nous trouvons intéressant de se concentrer sur deux points. Le premier point est d'étudier statistiquement la relation de dépendance entre les deux modalités sur de grands échantillons. Le deuxième est d'établir un lien entre cette relation est les variations du visage. Il s'agit d'étudier la stabilité de la relation selon les variations (éclairage, expression et pose) constatées, et les modes éventuels de la variation de cette relation. Une bonne compréhension de ces deux points devrait permettre de mieux combiner les données des deux modalités de façon à bénéficier pleinement de leur fusion.

Bibliographie

[1] Abate, A.F., Nappi, M., Riccio, D., Sabatino, G. : 2D and 3D face recognition : A survey. Pattern Recognition Letters **28**(14), 1885 – 1906 (2007)

[2] Achermann, B., Jiang, X., Bunke, H. : Face recognition using range images. In : Virtual Systems and MultiMedia. Proceedings of the International Conference on, pp. 129–136. IEEE (1997)

[3] Ahlberg, J. : Candide-3 - an updated parameterised face. Tech. rep., No. LiTH-ISY-R-2326, Dept. of Electrical Engineering, Linköping University (2001)

[4] Ahonen, T., Hadid, A., Pietikainen, M. : Face recognition with local binary patterns. In : Computer Vision. European Conference on, *Lecture Notes in Computer Science*, vol. 3021, pp. 469–481. Springer Berlin Heidelberg (2004)

[5] Ahonen, T., Hadid, A., Pietikainen, M. : Face description with local binary patterns : Application to face recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on **28**(12), 2037–2041 (2006)

[6] Amor, B., Ardabilian, M., Chen, L. : Enhancing 3D face recognition by mimics segmentation. In : Intelligent Systems Design and Applications. Sixth International Conference on, vol. 3, pp. 150–155 (2006)

[7] Arca, S., Lanzarotti, R., Lipori, G. : Face recognition based on 2D and 3D features. In : B. Apolloni, R. Howlett, L. Jain (eds.) Knowledge-Based Intelligent Information and Engineering Systems, *Lecture Notes in Computer Science*, vol. 4692, pp. 455–462. Springer Berlin Heidelberg (2007)

[8] Atick, J.J., Griffin, P.A., Redlich, A.N. : Statistical approach to shape from shading : Reconstruction of three-dimensional face surfaces from single two-dimensional images. Neural Computation **8**(6), 1321–1340 (1996)

[9] Baker, S., Matthews, I. : Equivalence and efficiency of image alignment algorithms. In : Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 1090 – 1097 (2001)

[10] Bardsley, D.J., Bai, L. : 3D surface reconstruction and recognition. In : Proceeding of SPIE, vol. 6539, pp. 653,906–1 (2007)

[11] Bartlett, M.S. : Independent component representations for face recognition. In : Face Image Analysis by Unsupervised Learning, pp. 39–67. Springer (2001)

[12] Bartlett, M.S., Movellan, J.R., Sejnowski, T.J. : Face recognition by independent component analysis. Neural Networks, IEEE Transactions on **13**(6), 1450–1464 (2002)

[13] Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J. : Eigenfaces vs. fisherfaces : Recognition using class specific linear projection. Pattern Analysis and Machine Intelligence, IEEE Transactions on **19**(7), 711–720 (1997)

[14] BenAbdelkader, C., Griffin, P.A. : Comparing and combining depth and texture cues for face recognition. Image Vision Computing. **23**(3), 339–352 (2005)

[15] Berretti, S., Del Bimbo, A., Pala, P. : 3D face recognition using isogeodesic stripes. Pattern Analysis and Machine Intelligence, IEEE Transactions on **32**(12), 2162 –2177 (2010)

[16] Besl, P.J., McKay, N.D. : Method for registration of 3-d shapes. In : Robotics-DL tentative, pp. 586–606. International Society for Optics and Photonics (1992)

[17] Beumier, C., Acheroy, M. : Face verification from 3D and grey level clues. Pattern Recognition Letters **22**(12), 1321 – 1329 (2001). Selected Papers from the 11th Portuguese Conference on Pattern Recognition - {RECPAD2000}

[18] Beveridge, J.R., Bolme, D., Draper, B.A., Teixeira, M. : The csu face identification evaluation system. Machine vision and applications **16**(2), 128–138 (2005)

[19] Beveridge, J.R., She, K., Draper, B.A., Givens, G.H. : A nonparametric statistical comparison of principal component and linear discriminant subspaces for face recognition. In : Computer Vision and Pattern Recognition. Proceedings of the IEEE Computer Society Conference on, vol. 1, pp. I–535. IEEE (2001)

[20] Birchfield, S., Tomasi, C. : Depth discontinuities by pixel-to-pixel stereo. International Journal of Computer Vision **35**(3), 269–293 (1999)

[21] Blais, F. : Review of 20 years of range sensor development. Journal of Electronic Imaging **13**(1) (2004)

[22] Blanz, V., Vetter, T. : A morphable model for the synthesis of 3D faces. In : Proceedings of the 26th annual conference on Computer graphics and interactive techniques, pp. 187–194. ACM Press/Addison-Wesley Publishing Co. (1999)

[23] Blanz, V., Vetter, T. : Face recognition based on fitting a 3D morphable model. Pattern Analysis and Machine Intelligence, IEEE Transactions on **25**(9), 1063–1074 (2003)

[24] Bledsoe, W. : Man-machine facial recognition. Rep. Panoramic Research Inc. **22** (1966)

[25] Boser, B.E., Guyon, I.M., Vapnik, V.N. : A training algorithm for optimal margin classifiers. In : Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory, pp. 144–152. ACM Press (1992)

[26] Bowyer, K.W., Chang, K., Flynn, P. : A survey of approaches and challenges in 3D and multi-modal 3D + 2D face recognition. Computer Vision and Image Understanding **101**(1), 1–15 (2006)

[27] Brand, M. : Morphable 3D models from video. In : Computer Vision and Pattern Recognition. Proceedings of the IEEE Computer Society Conference on, vol. 2, pp. II–456–II–463 vol.2 (2001)

[28] Bronstein, A.M., Bronstein, M.M., Kimmel, R. : Three-dimensional face recognition. International Journal of Computer Vision **64**(1), 5–30 (2005)

[29] Brunelli, R., Poggio, T. : Face recognition : Features versus templates. Pattern Analysis and Machine Intelligence, IEEE Transactions on **15**(10), 1042–1052 (1993)

[30] Castanedo, F. : A review of data fusion techniques. The Scientific World Journal **2013** (2013)

[31] Chang, K., Bowyer, K., Flynn, P. : Face recognition using 2D and 3D facial data. In : ACM Workshop on Multimodal User Authentication, pp. 25–32. Citeseer (2003)

[32] Chien, J.T., Liao, C.P. : Maximum confidence hidden markov modeling for face recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on **30**(4), 606–616 (2008)

[33] Choi, J., Medioni, G., Lin, Y., Silva, L., Regina, O., Pamplona, M., Faltemier, T. : 3D face reconstruction using a single or multiple views. In : Pattern Recognition. 20th International Conference on, pp. 3959 –3962 (2010)

[34] Chow, C., Yuen, S. : Recovering shape by shading and stereo under lambertian shading model. International journal of computer vision **85**(1), 58–100 (2009)

[35] Chowdhury, A., Chellappa, R., Krishnamurthy, S., Vo, T. : 3D face reconstruction from video using a generic model. In : Multimedia and Expo. Proceedings. 2002 IEEE International Conference on, vol. 1, pp. 449–452 vol.1 (2002)

[36] Chua, C.S., Jarvis, R. : Point signatures : A new representation for 3D object recognition. International Journal of Computer Vision **25**(1), 63–85 (1997)

[37] Colbry, D., Stockman, G. : The 3DID face alignment system for verifying identity. Image and Vision Computing **27**(8), 1121–1133 (2009)

[38] Cootes, T., Edwards, G., Taylor, C. : Active appearance models. Pattern Analysis and Machine Intelligence, IEEE Transactions on **23**(6), 681–685 (2001)

[39] Cootes, T., Walker, K., Taylor, C. : View-based active appearance models. In : Automatic Face and Gesture Recognition. Fourth IEEE International Conference on, pp. 227–232 (2000)

[40] Cootes, T.F., Taylor, C.J., et al. : Statistical models of appearance for computer vision. Imaging Science and Biomedical Engineering, University of Manchester, Manchester M13 9PT, UK March **8** (2004)

[41] Cryer, J., Tsai, P., Shah, M. : Integration of shape from shading and stereo. Pattern recognition **28**(7), 1033–1043 (1995)

[42] Danisman, T., Bilasco, I.M., Ihaddadene, N., Djeraba, C. : Automatic facial feature detection for facial expression recognition. In : VISAPP (2), pp. 407–412 (2010)

[43] Draper, B.A., Yambor, W.S., Beveridge, J.R. : Analyzing pca-based face recognition algorithms : Eigenvector selection and distance measures. Empirical Evaluation Methods in Computer Vision, Singapore (2002)

[44] Drira, H., Amor, B.B., Srivastava, A., Daoudi, M., Slama, R. : 3D face recognition under expressions, occlusions, and pose variations. Pattern Analysis and Machine Intelligence, IEEE Transactions on **35**(9), 2270–2283 (2013)

[45] Durrant-Whyte, H.F. : Sensor models and multisensor integration. International Journal of Robotics Research **7**(6), 97–113 (1988)

[46] Edwards, G.J., Cootes, T.F., Taylor, C.J. : Face recognition using active appearance models. In : Computer Vision, European Conference on, pp. 581–595. Springer (1998)

[47] Faltemier, T., Bowyer, K., Flynn, P. : A region ensemble for 3-d face recognition. Information Forensics and Security, IEEE Transactions on **3**(1), 62 –73 (2008)

[48] Fanany, M.I., Ohno, M., Kumazawa, I. : Face reconstruction from shading using smooth projected polygon representation nn. In : Proceedings of the 15th International Conference on Vision Interface, Calgary, Canada, pp. 308–313 (2002)

[49] Fisher, R.A. : The use of multiple measurements in taxonomic problems. Annals of eugenics **7**(2), 179–188 (1936)

[50] Fortuna, J., Martinez, A. : Rigid structure from motion from a blind source separation perspective. International Journal of Computer Vision **88**(3), 404–424 (2010)

[51] Freund, Y., Schapire, R.E. : A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences **55**(1), 119 – 139 (1997). URL `http://www.sciencedirect.com/science/article/pii/S002200009791504X`

[52] Gao, Y., Leung, M.K.H. : Face recognition using line edge map. Pattern Analysis and Machine Intelligence, IEEE Transactions on **24**(6), 764–779 (2002)

[53] Gokberk, B., Akarun, L. : Comparative analysis of decision-level fusion algorithms for 3D face recognition. In : Pattern Recognition. 18th International Conference on, vol. 3, pp. 1018–1021 (2006)

[54] Gökberk, B., Salah, A.A., Akarun, L. : Rank-based decision fusion for 3D shape-based face recognition. In : Audio-and Video-Based Biometric Person Authentication, pp. 1019–1028. Springer (2005)

[55] Gross, R., Baker, S., Matthews, I., Kanade, T. : Face recognition across pose and illumination. In : IN HANDBOOK OF FACE RECOGNITION. Springer-Verlag (2004)

[56] Gupta, S., Castleman, K., Markey, M., Bovik, A. : Texas 3D face recognition database. In : Image Analysis and Interpretation. IEEE Southwest Symposium on, pp. 97–100. IEEE (2010)

[57] Hagan, M.T., Demuth, H.B., Beale, M.H., et al. : Neural network design. Pws Pub. Boston (1996)

[58] Hajati, F., Raie, A.A., Gao, Y. : 2.5D face recognition using patch geodesic moments. Pattern Recognition **45**(3), 969 – 982 (2012)

[59] Hall, D., Llinas, J.A. : An introduction to multisensor data fusion. Proceedings of the IEEE **85**(1), 6–23 (1997)

[60] Heseltine, T., Pears, N., Austin, J. : Three-dimensional face recognition : An eigensurface approach. In : Image Processing. International Conference on, vol. 2, pp. 1421–1424. IEEE (2004)

[61] Hesher, C., Srivastava, A., Erlebacher, G. : A novel technique for face recognition using range imaging. In : Signal processing and its applications. Proceedings of the seventh international symposium on, vol. 2, pp. 201–204. IEEE (2003)

[62] Hirschmuller, H. : Improvements in real-time correlation-based stereo vision. In : Stereo and Multi-Baseline Vision. Proceedings. IEEE Workshop on, pp. 141–148. IEEE (2001)

[63] Horn, B.K. : Shape from shading : A method for obtaining the shape of a smooth opaque object from one view (1970)

[64] Huang, D., Ardabilian, M., Wang, Y., Chen, L. : Automatic asymmetric 3D-2D face recognition. In : Pattern Recognition. 20th International Conference on, pp. 1225–1228 (2010)

[65] Huang, D., Ardabilian, M., Wang, Y., Chen, L. : 3-d face recognition using elbp-based facial description and local feature hybrid matching. Information Forensics and Security, IEEE Transactions on **7**(5), 1551–1565 (2012)

[66] Huang, D., Ouji, K., Ardabilian, M., Wang, Y., Chen, L. : 3D face recognition based on local shape patterns and sparse representation classifier. Advances in Multimedia Modeling pp. 206–216 (2011)

[67] Huang, D., Shan, C., Ardabilian, M., Wang, Y., Chen, L. : Local binary patterns and its application to facial image analysis : A survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C **41**(6), 765–781 (2011)

[68] Huang, D., Zhang, G., Ardabilian, M., Wang, Y., Chen, L. : 3D face recognition using distinctiveness enhanced facial representations and local feature hybrid matching. In : Biometrics : Theory Applications and Systems. Fourth IEEE International Conference on, pp. 1–7 (2010)

[69] Huang, Y., Wang, Y., Tan, T. : Combining statistics of geometrical and correlative features for 3D face recognition. In : Proceedings of the British Machine Vision Conference, pp. 879–888 (2006)

[70] Husken, M., Brauckmann, M., Gehlen, S., Von der Malsburg, C. : Strategies and benefits of fusion of 2D and 3D face recognition. In : Computer Vision and Pattern Recognition-Workshops. IEEE Computer Society Conference on, pp. 174–174. IEEE (2005)

[71] Jahanbin, S., Choi, H., Bovik, A. : Passive multimodal 2-d+3-d face recognition using gabor features and landmark distances. Information Forensics and Security, IEEE Transactions on **6**(4), 1287–1304 (2011)

[72] Jain, A.K., Li, S.Z. : Handbook of face recognition. Springer (2005)

[73] Jebara, T.S., Pentland, A. : Parametrized structure from motion for 3D adaptive feedback tracking of faces. In : Proceedings of Computer Vision and Pattern Recognition, pp. 144–150 (1997)

[74] Kakadiaris, I., Passalis, G., Toderici, G., Murtuza, M., Lu, Y., Karampatziakis, N., Theoharis, T. : Three-dimensional face recognition in the presence of facial expressions : An annotated deformable model approach. Pattern Analysis and Machine Intelligence, IEEE Transactions on **29**(4), 640 –649 (2007)

[75] Kanade, T. : Computer recognition of human faces. Birkhäuser (1977)

[76] Kanade, T., Morris, D.D. : Factorization methods for structure from motion. philosophical transactions of the royal society of London, series a **356**, 1153–1173 (2001)

[77] Kano, H., Ghosh, B., Kanai, H. : Single camera based motion and shape estimation using extended kalman filtering. Mathematical and Computer Modelling **34**, 511 – 525 (2001)

[78] Kellogg, T., Truesdale, R., Kellogg, ., Osterman, L., DW, M., Brady, H., Martin, H., Webb, P. : A computer algorithm for reconstructing a scene from two projections. Nature **293**, 133 (1981)

[79] Kim, T.K., Kim, H., Hwang, W., Kee, S.C., Kittler, J. : Independent component analysis in a facial local residue space. In : Computer Vision and Pattern Recognition. Proceedings of the IEEE Computer Society Conference on, vol. 1, pp. I–579. IEEE (2003)

[80] Kolmogorov, V., Zabih, R. : Multi-camera scene reconstruction via graph cuts. In : Computer Vision, European Conference on, pp. 82–96 (2003)

[81] Koo, H.S., Lam, K.M. : Recovering the 3D shape and poses of face images based on the similarity transform. Pattern Recognition Letters **29**(6), 712 – 723 (2008)

[82] Kusuma, G., Chua, C.S. : Image level fusion method for multimodal 2D+3D face recognition. In : A. Campilho, M. Kamel (eds.) Image Analysis and Recognition, *Lecture Notes in Computer Science*, vol. 5112, pp. 984–992. Springer Berlin Heidelberg (2008)

[83] Lanitis, A., Taylor, C., Cootes, T. : A unified approach to coding and interpreting face images. In : Computer Vision. Proceedings of the Fifth International Conference on, pp. 368–373. IEEE (1995)

[84] Lanitis, A., Taylor, C.J., Cootes, T.F. : Automatic interpretation and coding of face images using flexible models. Pattern Analysis and Machine Intelligence, IEEE Transactions on **19**(7), 743–756 (1997)

[85] Le, V., Tang, H., Cao, L., Huang, T. : Accurate and efficient reconstruction of 3D faces from stereo images. In : Image Processing. 17th IEEE International Conference on, pp. 4265 –4268 (2010)

[86] Lee, J., Moghaddam, B., Pfister, H., Machiraju, R., Lee, J., Moghaddam, B., Pfister, H., Machiraju, R. : Silhouette-based 3D face shape recovery. In : Graphics Interface, pp. 21–30 (2003)

[87] Lee, J.C., Milios, E. : Matching range images of human faces. In : Computer Vision. Proceedings of the third International Conference on, pp. 722–726. IEEE (1990)

[88] Lee, M., Choi, C.H. : Facial shape recovery from a single image with an arbitrary directional light using linearly independent representation. In : Advances in Visual Computing, pp. 740–749. Springer (2009)

[89] Lee, Y., Song, H., Yang, U., Shin, H., Sohn, K. : Local feature based 3D face recognition. In : Audio-and Video-Based Biometric Person Authentication, pp. 909–918. Springer (2005)

[90] Lengagne, R., Fua, P., Monga, O. : 3D stereo reconstruction of human faces driven by differential constraints. Image and Vision Computing **18**(4), 337–343 (2000)

[91] Li, S., Zhao, C., Ao, M., Lei, Z. : Learning to fuse 3D+2D based face recognition at both feature and decision levels. In : W. Zhao, S. Gong, X. Tang (eds.) Analysis and Modelling of Faces and Gestures, *Lecture Notes in Computer Science*, vol. 3723, pp. 44–54. Springer Berlin Heidelberg (2005)

[92] Lin, W.Y., Chen, M.Y. : A novel framework for automatic 3D face recognition using quality assessment. Multimedia Tools and Applications pp. 1–17 (2012)

[93] Liu, D.H., Lam, K.M., Shen, L.S. : Illumination invariant face recognition. Pattern Recognition **38**(10), 1705 – 1716 (2005). URL `http://www.sciencedirect.com/science/article/pii/S0031320305001378`

[94] Liu, L., Zhao, L., Long, Y., Kuang, G., Fieguth, P. : Extended local binary patterns for texture classification. Image and Vision Computing **30**(2), 86 – 99 (2012)

[95] Luong, Q.T., Deriche, R., Faugeras, O., Papadopoulo, T. : On determining the fundamental matrix : analysis of different methods and experimental results. Rapport de recherche RR-1894, INRIA (1993)

[96] Mallick, S.P., Trivedi, M. : Parametric face modeling and affect synthesis. In : Proceedings of the International Conference on Multimedia and Expo - Volume 2, ICME '03, pp. 225–228. IEEE Computer Society, Washington, DC, USA (2003)

[97] Martiriggiano, T., Leo, M., D'Orazio, T., Distante, A. : Face recognition by kernel independent component analysis. In : M. Ali, F. Esposito (eds.) Innovations in Applied Artificial Intelligence, *Lecture Notes in Computer Science*, vol. 3533, pp. 55–58. Springer Berlin Heidelberg (2005)

[98] Mian, A., Pears, N. : 3D face recognition. In : N. Pears, Y. Liu, P. Bunting (eds.) 3D Imaging, Analysis and Applications, pp. 311–366. Springer London (2012)

[99] Mian, A.S., Bennamoun, M., Owens, R. : An efficient multimodal 2D-3D hybrid approach to automatic face recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on **29**(11), 1927–1943 (2007)

[100] Mika, S., Ratsch, G., Weston, J., Scholkopf, B., Mullers, K. : Fisher discriminant analysis with kernels. In : Neural Networks for Signal Processing IX. Proceedings of the IEEE Signal Processing Society Workshop., pp. 41–48. IEEE (1999)

[101] Milborrow, S., Nicolls, F. : Locating facial features with an extended active shape model. Computer Vision. European Conference on pp. 504–513 (2008)

[102] Mitra, S., Parua, S., Das, A., Mazumdar, D. : A novel data mining approach for performance improvement of ebgm based face recognition engine to handle large database. In : Advances in Computer Science and Information Technology, pp. 532–541. Springer (2011)

[103] Moreno, A.B., Sánchez, A., Vélez, J.F., Díaz, F.J. : Face recognition using 3D surface-extracted descriptors. In : Irish Machine Vision and Image. Processing Conference on. Citeseer (2003)

[104] Ojala, T., Pietikäinen, M., Mäenpää, T. : Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence. **24**(7), 971–987 (2002)

[105] Ojala, T., Valkealahti, K., Oja, E., Pietikäinen, M. : Texture discrimination with multidimensional distributions of signed gray-level differences. Pattern Recognition **34**(3), 727 – 739 (2001)

[106] Papatheodorou, T., Rueckert, D. : Evaluation of automatic 4d face recognition using surface and texture registration. In : Automatic Face and Gesture Recognition. Proceedings of the sixth IEEE International Conference on, pp. 321–326. IEEE (2004)

[107] Park, U., Jain, A.K. : 3D face reconstruction from stereo video. In : Computer and Robot Vision. The 3rd Canadian Conference on, p. 41 (2006)

[108] Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T. : A 3D face model for pose and illumination invariant face recognition. In : Advanced Video and Signal Based Surveillance. Sixth IEEE International Conference on, pp. 296–301. IEEE (2009)

[109] Pentland, A., Moghaddam, B., Starner, T. : View-based and modular eigenspaces for face recognition. In : Computer Vision and Pattern Recognition. Proceedings of the IEEE Computer Society Conference on, pp. 84–91. IEEE (1994)

[110] Phillips, P.J., Flynn, P.J., Scruggs, T., Bowyer, K.W., Chang, J., Hoffman, K., Marques, J., Min, J., Worek, W. : Overview of the face recognition grand challenge. In : Computer vision and pattern recognition. IEEE computer society conference on, vol. 1, pp. 947–954. IEEE (2005)

[111] Phillips, P.J., Grother, P., Micheals, R., Blackburn, D.M., Tabassi, E., Bone, M. : Face recognition vendor test 2002. In : Analysis and Modeling of Faces and Gestures. IEEE International Workshop on, p. 44. IEEE (2003)

[112] Phillips, P.J., Moon, H., Rizvi, S.A., Rauss, P.J. : The feret evaluation methodology for face-recognition algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(10), 1090–1104 (2000)

[113] Price, J.R., Gee, T.F. : Face recognition using direct, weighted linear discriminant analysis and modular subspaces. Pattern Recognition **38**(2), 209–219 (2005)

[114] Romdhani, S., Vetter, T. : Efficient, robust and accurate fitting of a 3D morphable model. In : Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2, ICCV '03, pp. 59–. IEEE Computer Society, Washington, DC, USA (2003)

[115] Salvi, J., Pages, J., Batlle, J. : Pattern codification strategies in structured light systems. Pattern Recognition **37**(4), 827–849 (2004)

[116] Samaria, F., Young, S. : Hmm-based architecture for face identification. Image and vision computing **12**(8), 537–543 (1994)

[117] Sanderson, C., Paliwal, K. : Information fusion and person verification using speech & face information (2002)

[118] Savran, A., Alyüz, N., Dibeklioğlu, H., Çeliktutan, O., Gökberk, B., Sankur, B., Akarun, L. : Bosphorus database for 3D face analysis. In : Biometrics and Identity Management, pp. 47–56. Springer (2008)

[119] Scharstein, D., Szeliski, R. : A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International Journal of Computer Vision. **47**, 7–42 (2002)

[120] Scharstein, D., Szeliski, R. : High-accuracy stereo depth maps using structured light. In : Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on, vol. 1, pp. I–195–I–202 vol.1 (2003)

[121] Sirovich, L., Kirby, M. : Low-dimensional procedure for the characterization of human faces. JOSA A **4**(3), 519–524 (1987)

[122] Smith, W.A., Hancock, E.R. : Facial shape-from-shading and recognition using principal geodesic analysis and robust statistics. International Journal of Computer Vision **76**(1), 71–91 (2008)

[123] Soltana, W.B., Huang, D., Ardabilian, M., Chen, L., Amar, C.B. : Comparison of 2D/3D features and their adaptive score level fusion for 3D face recognition. 3D Data Processing, Visualization and Transmission (3DPVT) (2010)

[124] Spreeuwers, L. : Fast and accurate 3D face recognition. International Journal of Computer Vision **93**, 389–414 (2011)

[125] Sun, J., Zheng, N.N., Shum, H.Y. : Stereo matching using belief propagation. Pattern Analysis and Machine Intelligence, IEEE Transactions on **25**(7), 787 – 800 (2003)

[126] Sun, T.H., Chen, M., Lo, S., Tien, F.C. : Face recognition using 2D and disparity eigenface. Expert Systems with Applications **33**(2), 265 – 273 (2007)

[127] Sun, Z., Lam, K.M. : Depth estimation of face images based on the constrained ica model. Information Forensics and Security, IEEE Transactions on **6**(2), 360–370 (2011)

[128] Sun, Z.L., Lam, K.M., Gao, Q. : Depth estimation of face images using the nonlinear least-squares model. IEEE Transactions on Image Processing **22**(1), 17–30 (2013)

[129] Takacs, B. : Comparing face images using the modified hausdorff distance. Pattern Recognition **31**(12), 1873–1881 (1998)

[130] Tan, X., Triggs, B. : Enhanced local texture feature sets for face recognition under difficult lighting conditions. Image Processing, IEEE Transactions on **19**(6), 1635–1650 (2010)

[131] Tanaka, H.T., Ikeda, M., Chiaki, H. : Curvature-based face surface recognition using spherical correlation. principal directions for curved object recognition. In : Automatic Face and Gesture Recognition. Proceedings of the third IEEE International Conference on, pp. 372–377. IEEE (1998)

[132] Tang, H., Sun, Y., Yin, B., Ge, Y. : Expression-robust 3D face recognition using lbp representation. In : Multimedia and Expo. IEEE International Conference on, pp. 334–339. IEEE (2010)

[133] Tang, H., Yin, B., Sun, Y., Hu, Y. : 3D face recognition using local binary patterns. Signal Processing **93**(8), 2190 – 2198 (2013)

[134] Thalmann, N., Kalra, P., Escher, M. : Face to virtual face. Proceedings of the IEEE **86**(5), 870–883 (1998)

[135] Torresani, L., Hertzmann, A., Bregler, C. : Nonrigid structure-from-motion : Estimating shape and motion with hierarchical priors. IEEE Transactions on Pattern Analysis and Machine Intelligence. **30**(5), 878–892 (2008)

[136] Trucco, E., Verri, A. : Introductory Techniques for 3-D Computer Vision. Prentice Hall PTR, Upper Saddle River, NJ, USA (1998)

[137] Tsalakanidou, F., Malassiotis, S., Strintzis, M. : Face localization and authentication using color and depth images. Image Processing, IEEE Transactions on **14**(2), 152–168 (2005)

[138] Tsalakanidou, F., Tzovaras, D., Strintzis, M. : Use of depth and colour eigenfaces for face recognition. Pattern Recognition Letters **24**(9), 1427–1435 (2003)

[139] Turk, M.A., Pentland, A.P. : Face recognition using eigenfaces. In : Computer Vision and Pattern Recognition. Proceedings of the IEEE Computer Society Conference on, pp. 586–591. IEEE (1991)

[140] Viola, P., Jones, M. : Robust real-time object detection. In : International Journal of Computer Vision (2001)

[141] Wang, J.G., Lim, E., Chen, X., Venkateswarlu, R. : Real-time stereo face recognition by fusing appearance and depth fisherfaces. The Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology **49**(3), 409–423 (2007)

[142] Wang, L., Ding, L., Ding, X., Fang, C. : Improved 3D assisted pose-invariant face recognition. In : Acoustics, Speech and Signal Processing. IEEE International Conference on, pp. 889–892. IEEE (2009)

[143] Wang, S., Zhang, L., Samaras, D. : Face reconstruction across different poses and arbitrary illumination conditions. In : T. Kanade, A. Jain, N. Ratha (eds.) Audio- and Video-Based Biometric Person Authentication, *Lecture Notes in Computer Science*, vol. 3546, pp. 91–101. Springer Berlin Heidelberg (2005)

[144] Wang, S.F., Lai, S.H. : Reconstructing 3D face model with associated expression deformation from a single face image via constructing a low-dimensional expression deformation manifold. Pattern Analysis and Machine Intelligence, IEEE Transactions on **33**(10), 2115 –2121 (2011)

[145] Wang, X., Ruan, Q., Ming, Y. : 3D face recognition using corresponding point direction measure and depth local features. In : Signal Processing. IEEE 10th International Conference on, pp. 86–89 (2010)

[146] Wang, Y., Chua, C.S., Ho, Y.K. : Facial feature detection and face recognition from 2D and 3D images. Pattern Recognition Letters **23**(10), 1191–1202 (2002)

[147] Wang, Y., Liu, J., Tang, X. : Robust 3D face recognition by local shape difference boosting. Pattern Analysis and Machine Intelligence, IEEE Transactions on **32**(10), 1858 –1870 (2010)

[148] Wang, Y., Liu, J., Tang, X. : Robust 3D face recognition by local shape difference boosting. Pattern Analysis and Machine Intelligence, IEEE Transactions on **32**(10), 1858–1870 (2010)

[149] Weng, J., Huang, T., Ahuja, N. : Motion and structure from line correspondences ; closed-form solution, uniqueness, and optimization. Pattern Analysis and Machine Intelligence, IEEE Transactions on **14**(3), 318–336 (1992)

[150] Weyrauch, B., Heisele, B., Huang, J., Blanz, V. : Component-based face recognition with 3D morphable models. In : Computer Vision and Pattern Recognition Workshop. Conference on, pp. 85–85. IEEE (2004)

[151] Wiskott, L., Fellous, J.M., Kuiger, N., von der Malsburg, C. : Face recognition by elastic bunch graph matching. Pattern Analysis and Machine Intelligence, IEEE Transactions on **19**(7), 775–779 (1997)

[152] Wiskott, L., von der Malsburg, C. : Recognizing faces by dynamic link matching. Neuroimage **4**(3), S14–S18 (1996)

[153] Wright, J., Hua, G. : Implicit elastic matching with random projections for pose-variant face recognition. In : Computer Vision and Pattern Recognition. IEEE Conference on, pp. 1502–1509. IEEE (2009)

[154] Xiao, J., Baker, S., Matthews, I., Kanade, T. : Real-time combined 2D+3D active appearance models. In : Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 535 – 542 (2004)

[155] Xiong, P., Huang, L., Liu, C. : Real-time 3D face recognition with the integration of depth and intensity images. In : M. Kamel, A. Campilho (eds.) Image Analysis and Recognition, *Lecture Notes in Computer Science*, vol. 6754, pp. 222–232. Springer Berlin Heidelberg (2011)

[156] Xu, C., Li, S., Tan, T., Quan, L. : Automatic 3D face recognition from depth and intensity gabor features. Pattern Recognition **42**(9), 1895 – 1905 (2009)

[157] Xu, L., Krzyzak, A., Suen, C. : Methods of combining multiple classifiers and their applications to handwriting recognition. Systems, Man and Cybernetics, IEEE Transactions on **22**(3), 418–435 (1992)

[158] Yan, P., Bowyer, K.W. : A fast algorithm for icp-based 3D shape biometrics. Computer Vision and Image Understanding **107**(3), 195 – 202 (2007)

[159] Yang, M.H. : Kernel eigenfaces vs. kernel fisherfaces : Face recognition using kernel methods. In : Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition, p. 215. Washington, DC (2002)

[160] Zhang, L., Chu, R., Xiang, S., Liao, S., Li, S. : Face detection based on multi-block lbp representation. In : S.W. Lee, S. Li (eds.) Advances in Biometrics, *Lecture Notes in Computer Science*, vol. 4642, pp. 11–18. Springer Berlin Heidelberg (2007)

[161] Zhang, R., Tsai, P.S., Cryer, J., Shah, M. : Shape-from-shading : a survey. Pattern Analysis and Machine Intelligence, IEEE Transactions on **21**(8), 690–706 (1999)

[162] Zhao, G., Pietikainen, M. : Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence. **29**(6), 915–928 (2007)

[163] Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A. : Face recognition : A literature survey. Acm Computing Surveys **35**(4), 399–458 (2003)

[164] Zheng, Y., Chang, J., Zheng, Z., Wang, Z. : 3D face reconstruction from stereo : A model based approach. In : IEEE International Conference on Image Processing., pp. III –65 –III –68 (2007)

[165] Zhou, S., Chellappa, R. : Illuminating light field : Image-based face recognition across illuminations and poses. In : Automatic Face and Gesture Recognition. Proceedings of the sixth IEEE International Conference on, pp. 229–234. IEEE (2004)

[166] Zucchelli, M., Santos-Victor, J., Christensen, H.I. : Constrained structure and motion estimation from optical flow. In : International Conference on Pattern Recognition., pp. 339–342 (2002)